Padoa-Schioppa, Tommaso (1940–2010)

Lorenzo Bini Smaghi

Abstract

Tommaso Padoa-Schioppa was a distinguished central banker and economist, a key player in the creation of Europe's single currency, a well-respected figure in international monetary policy-making and a former economy and finance minister of Italy.

Keywords

Central bank; Economic integration; European Union; Financial stability; Single currency

JEL Classifications

B31; E5; E58; E52; F15

Tommaso Padoa-Schioppa

Introduction

Born in Belluno (Italy) on 23 July 1940, Tommaso Padoa-Schioppa was an economist, a policy-maker and a citizen of Europe. A deep and complex personality, he pursued a Kantian quest for European unity, a mission to which he dedicated his professional and private life. He died in Rome on 18 December 2010.

The Economist and Policy-Maker

Padoa-Schioppa joined the Banca d'Italia in 1968. Italy's central bank, with its tradition of technical expertise and high standards of public service, was the obvious place to go for an economist wishing to serve his country. As a young economist working in the bank's research department, Padoa-Schioppa was keen to combine doing and understanding.

In the late 1960s the Bretton Woods system, which had given Italy a solid monetary framework, was showing signs of wear; the growth 'miracle' of the Italian economy was starting to lose momentum. The country had problems shaping an economic policy appropriate for those changing times: budgetary policy was slipping out of control and was set to turn into a burden for decades afterwards. Monetary policy, which was about to lose its exchange rate anchor to the dollar, was casting around for an alternative. Padoa-Schioppa saw the efforts to establish a

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

European monetary order not only as a chance to make decisive progress towards a unified Europe but also as an opportunity to help Italy overcome its economic problems in a sound and stability-oriented manner.

Padoa-Schioppa was influenced by some of the greatest post-war economists. Among them was Mundell, who, together with Hayek, discussed the ‘denationalisation’ of currencies. In his *Lectio Doctoralis* when he was awarded a *Laurea Honoris Causa* in International Economics of Trade and Currency Markets, in Trieste on 19 November 1999, Padoa-Schioppa argued that Mundell had sown the seeds which would complete the Treaty of Rome and established the theoretical foundations of the Maastricht Treaty. Triffin, with his book *Gold and the Dollar Crisis* (1960), also strongly influenced Padoa-Schioppa, as argued below.

Against this theoretical background, Padoa-Schioppa supported Italy’s membership of the European Monetary System (EMS), in spite of those who argued (with some reason) that Italy would have difficulty in complying with the exchange rate constraints of the EMS. He was, of course, aware of this, yet he did not want such an endeavour to start without Italian participation. Moreover, he believed that the so-called ‘monetarist’ view – whereby decisions taken on the exchange rate have effects on the conduct of monetary policy and on the economy – reflected the reality better than the so-called ‘economist’ view, according to which exchange rate constraints and a common currency could only be the climax of a perfect convergence of monetary policy and economic conditions.

These same ideas informed his action when he served at the European Commission from 1979 to 1983, the early years of the European Monetary System. There again, he faced an apparent trade-off between national and European forces and between monetary stability and a European monetary project. All his efforts as an economist and policy-maker were devoted to resolving these tensions and taking them in one direction – towards a European monetary endeavour that would bring monetary stability.

He took inspiration from the impossible trinity proposition, a corollary of the Mundell–Fleming model, which stated that a group of countries cannot simultaneously keep a fixed exchange rate, pursue autonomous monetary policies and maintain full capital mobility. In pursuing any two of these goals, the third one must be necessarily given up. The accomplishment of the central goal set down in the Treaty of Rome – a European common market, implying free trade and free capital movement – would inevitably call for enhanced, and eventually fixed, exchange rate stability across member countries. This is because exchange rate volatility, coupled with the systematic devaluation of a group of currencies against other currencies, would eventually impair the very functioning of the common market. Indeed, the whole history of European monetary cooperation between the collapse of Bretton Woods and the early 1990s is marked by attempts to maintain fixed exchange rates. These attempts were periodically punctuated by major tensions between national currencies (and consequently between Member States) each time a country was forced (or tempted) to devalue or even abandon the common fixed exchange rate arrangement. In this context, the economically most coherent way to reconcile intra-regional free trade and free capital movement with fixed exchange rates would be for European countries to adopt a single monetary policy and therefore a single currency floating against the rest of the world. In fact, Padoa-Schioppa preferred to talk about an ‘inconsistent quartet’ (*quartetto inconciliabile*), which included a fourth element, free trade, in order to take full account of the European context. In the early 1980s these ideas were still pioneering and visionary, but we all know how influential they have become in shaping the history of Europe, to the extent that they have transformed themselves from vision into reality. These ideas are contained in *Efficiency, Stability and Equity*, edited by Padoa-Schioppa and published in 1987.

Padoa-Schioppa followed the same course when he returned to the Banca d'Italia in 1984 and became Deputy Director-General, even if he had then to deal more directly with the continuing difficulties of Italian governance in respect of sound and stability-oriented macro policies.

While actively aiming to bring monetary stability to Italy and to move towards monetary unification in Europe, Padoa-Schioppa started working in payment systems – a relatively low-profile area. Before he took on responsibilities in this field at the Banca d'Italia, payment systems were not considered an independent (let alone a core) function of central banks. The activities carried out nowadays by payment system departments were largely non-existent at that time and performed mainly by IT departments and back offices.

Padoa-Schioppa provided the conceptual and economic rationale for a greater involvement by central banks in this area. A smooth functioning of payment systems is essential to ensure that money serves as a 'means of exchange', one of its key functions. Central banks came into being partly to ensure that this means of exchange remains adequate (i.e. confidence in the currency and its effective circulation). Central bank money is indeed universally recognised as the safest means of exchange and therefore as the ultimate way of discharging financial obligations. The smooth functioning of payment systems is also crucial for the effective conduct of monetary policy and financial stability.

Between 1985 and 1991, Padoa-Schioppa promoted and developed in Italy all the elements of an efficient and sound payment system. Under his guidance, the Banca d'Italia played a central role as an operator of settlement facilities, regulator and catalyst.

Padoa-Schioppa was the first person to use the word 'oversight' to describe the function to be conducted for payment systems, as opposed to '(banking) supervision'. He launched these activities before the Banca d'Italia was assigned formal specific statutory competence in this field. 'The function begets the organ and leads to the establishment of the legal basis', as he used to say, thereby acting as a catalyst. With hindsight, what Padoa-Schioppa did in Italy laid the foundations for the development of an analogous process at European level. His views on the role of central banks in payments systems can be found in particular in *The Euro and Its Central Bank*, published in 2004.

Padoa-Schioppa once again had a chance to be involved more actively in policy-making at European level when serving as secretary for the Delors Committee, which prepared the blueprint for the single currency in 1987–89. Although he was not a member of the Committee and, as such, was not supposed to intervene during the meetings, his personal relations with Jacques Delors and with Carlo Azeglio Ciampi, then Governor of the Banca d'Italia and a member of the Committee, as well as his role as 'rapporteur' allowed him to greatly influence the deliberations. It was an excellent opportunity to pursue his aspiration: a currency which was stable and which would significantly contribute to building the European Union.

In the 1989–91 negotiations that led to the Maastricht Treaty, Padoa-Schioppa fought to maintain the momentum towards monetary union which had been established by the Delors Committee. He feared that the decision to establish a temporary institution, the European Monetary Institute, instead of the European Central Bank, would be a critical mistake, also considering that a number of countries, including Italy, had ceased to conduct policies which were sufficiently stability-oriented.

The historic project of monetary unification came close to failure with the repeated exchange rate crises in the first half of the 1990s. But Europe's efforts eventually succeeded. In the 1993–97 preparatory work for the launch of the euro, Padoa-Schioppa had yet another chance to show his unique ability to build a strong conceptual, analytical and policy framework (the 'vision') on the one hand, and to implement a concrete set of measures and actions to turn the vision into reality on the other hand. His thoughts on the issue are summarized in *The Road to Monetary Union in Europe*, published in 1994, with the stimulating sub-title 'The Emperor, the Kings, and the Genies'.

In the field of financial stability, Padoa-Schioppa was convinced that in a single market, and even more so within a single currency area, the supervisory framework had to be as convergent as possible, entailing a transfer of supervisory responsibilities from national to EU level. He regarded the convergence of supervisory rules and of supervisory practices through enhanced cross-border cooperation as intermediate steps towards the final objective. It was he who coined the term 'EU supervisory rulebook' and, moreover, he was the founder of the Forum of European Securities Commissions, which promoted cooperation in the securities field. In his view, cross-border supervisory cooperation should be so strong and effective that the collective behaviour of supervisors would appear as a single effort. All these ideas started to take shape with the establishment of the Lamfalussy Committee, whose main purpose was to enhance supervisory convergence and cooperation. But his ideas have only been fully realised recently, with the setting-up in of three new European Supervisory Authorities (ESAs) as part of the European System of Financial Supervision, in January 2011.

Padoa-Schioppa was also convinced that central banks should play a key role in ensuring financial stability, as suggested in his 2004 book "Regulating Finance". Such a role should go beyond managing the risks incurred by individual financial institutions, a task performed by supervisors, and entail monitoring, controlling and managing the risks to the financial system as a whole (what is now commonly referred to as systemic risk). He essentially anticipated the issues which eventually came to the fore in the wide policy debate stemming from the financial crisis under the heading of macro-prudential supervision. This debate eventually led to the establishment of the European Systemic Risk Board.

He promoted these ideas in particular when he became Minister of Economy in the Italian government led by Romano Prodi (2006–08). As a Minister, Padoa-Schioppa found his country in a difficult situation. For a decade, its growth performance had been very disappointing, well below that of its main partners; Italy had also failed to maintain the progress in controlling public finances that it had achieved in the mid-1990s, in the run-up to the euro: the primary surplus, which had reached over five per cent of GDP, had been wiped out; public debt was on the rise again.

Padoa-Schioppa's three-pronged strategy – to pursue simultaneously the complementary objectives of stability, growth and social equity – soon became the manifesto of the government's economic policy. A programme of rigorous fiscal consolidation was in his view essential to revitalise the economy. He tried to instil a sense of urgency, in an often hostile political environment. In this field too he proved to be a reformer: he launched an ambitious spending review to overcome the purely incremental, history-dependent approach that dominated the political decisions on Italy's public budget. The objective was to reduce the quantity, and increase the quality, of public expenditure ("spend less, spend better"), by modifying the organisation of public administrations and their local branches, revising their structure according to the new needs, eliminating obsolete programmes and reconsidering the priorities, costs and ways of providing public services. The way forward was, in his words, "to promote excellence, by showing that there are already good practices in the country" so that it would become manifest that "we do not demand the impossible".

He largely succeeded in reversing the trend in public finances, attaining a significant reduction of the public deficit (from above four per cent of GDP in 2005 to below two per cent in 2007) and bringing the public debt-to-GDP ratio back on a downward path. This policy was unavoidable to get Italy out of its emergency situation; it was, for a country as indebted and under-capitalised as Italy, the only way forward to free up resources to invest in the future, in human and physical capital (infrastructures, education, research, environment) and to kick-start the Italian economy on a growth path again. His efforts concentrated primarily on the refinancing of infrastructural investments, an area where resources had completely dried up, and on reducing tax on labour and corporate income.

His brief spell as a Minister, coupled with the difficulties of a coalition government with a tiny majority in Parliament, prevented Padoa-Schioppa from exerting the more profound influence he would have been capable of. But his sense of direction and, especially, his determination to confront the problems of the Italian economy from a collective viewpoint and with a longer-run perspective than the country was used to, are still crucial points of reference for today's policy-makers.

In international relations, Padoa-Schioppa was convinced that the euro should gradually develop its role and policy at the global level: it owed this to its 'domestic constituency' as well as to the international community, as monetary unification in Europe could play an important role by enhancing global policy cooperation. In this area too, Padoa-Schioppa was a fervent advocate of the national central banks of the euro pooling their forces, acting as a system and playing a role consistent with the institutional mission and interests of the Eurosystem. His belief in international cooperation was founded on the conviction that there are, on the global agenda, many economic issues – including trade, finance, global imbalances, energy and the emergence of new global and regional players – whose scale and complexity call for policy efforts that go beyond national borders and make cooperation at global level desirable and feasible.

In the same spirit, Padoa-Schioppa considered it erroneous to believe that the euro and full national sovereignty were compatible. Indeed, he was critical of the notion of 'euro without a state'. In an increasingly globalised world political power cannot be concentrated in a single entity. This lesson is extraordinarily concrete and modern: Padoa-Schioppa was of the opinion that more integration is needed in domains ranging from fiscal policy to financial supervision.

Padoa-Schioppa belonged to that group of personalities who played a decisive role in shaping international cooperation. He contributed to formal and informal group discussions at global level, but also worked on a bilateral basis with countries, institutions, and people from Africa, Asia, Latin America and from the Mediterranean countries.

While Minister, he chaired the International Monetary and Financial Committee of the IMF, and played a key role in the process that eventually led to the reform of the quota and voting system of the IMF (finally adopted in April 2008). This brought the distribution of voting power more closely into line with the changing size of the member economies and resulted in better representation of the poorest countries.

In the performance of his international functions, Padoa-Schioppa was greatly influenced by J.M. Keynes and Robert Triffin. He argued that the 'Triffin dilemma' is still unresolved: if the main international standard and reserve currency is a national one there is an unresolvable incompatibility between domestic needs (in the short term) and global policy needs (in the long term). This incompatibility led, over time, to the collapse of the Bretton Woods system in the late 1960s and early 1970s. It was not resolved either under the US dollar-based non-system that took its place: a market-led system without a clear anchor. Over the last 10–15 years this has been referred to as Bretton Woods II. However, Padoa-Schioppa noted that 'exchange rates were determined by a bizarre combination of market behaviour and of policy actions vis-à-vis the dollar. Floating European currencies, including the euro, were at the mercy of a market prone to prolonged misalignments. Asian currencies were largely sheltered from the vagaries of the market and subjected to intense management by the national authorities'.

During his last years, Padoa-Schioppa became actively involved in an initiative of the Triffin International Foundation, which is based in Louvain-la-Neuve (Belgium). In his speech 'The Ghost of Bancor: the Economic Crisis and Global Monetary Disorder', given in Louvain in February 2010, he revisits the arguments proposed by Keynes in 1943 for a new international monetary system and eventually a world currency called Bancor.

The Legacy

Padoa-Schioppa's legacy as an economist, policymaker and citizen consists of at least six distinctive elements.

The first and most obvious is his faith in Europe. He saw Europe as a continent of peace, a 'gentle' force searching for its shared roots and a way to overcome problems that, if approached nationally, might retrigger the divisions and horrors of the past. Europe was always the focus, even when its representatives or actions did not deserve it. Padoa-Schioppa's Europe was an aspiration and an ideal, much bigger and longer lasting than the individuals that may at any given time symbolise it. His views can be found in particular in *Europe, a Civil Power*, published in 2004.

Padoa-Schioppa's approach to Europe was shaped by his views on federalism. He saw federalism as a constitutional system that owes a lot to the idea of *minimum government*, in the tradition of Locke and Tocqueville. According to his *Weltanschauung*, federalism would complement what he calls the 'horizontal division' of government functions – legislative, executive, judicial – with a 'vertical division', whereby government powers are distributed along all relevant levels. Such levels range from the village to the nation-state, from a regional arrangement to the whole world based on the principle of subsidiarity. This is the rule that 'the functions of higher levels of government should be as limited as possible and be subsidiary to those of lower levels' ('Economic Federalism and the European Union').

The second element of his legacy is that the future development of the Union can and will probably be built on the 'solidarité de fait' created by economic interdependence. Padoa-Schioppa was fond of reminding people that it was a single article in the US constitution – the one regulating inter-state commerce – that was used to lay the foundations for the federal government to have a significant role in US economic policies. But solidarity must go hand-in-hand with responsibility, and Europe's common destiny requires a reinforcement and a clarification of the rules of the game.

The third element in his approach to European and other issues was his faith in institutions. Only institutions (whether simple gatherings or forums, or formal debating and decision-making groups, up to structured organisations with physical premises) give continuity and strength to human efforts and aspirations that are, by nature, transient and mortal. 'Nothing is possible without humans, nothing is lasting without institutions', Jean Monnet wrote. The role of institutions in framing the operation of financial markets is well described in his BIS Per Jacobsson lecture given in June 2010, with the title 'Markets and government before, during and after the 2007–20xx crisis'.

The fourth key element is his approach to economics. There is no doubt that Padoa-Schioppa was an economist of the highest order: his early studies at MIT with Franco Modigliani, his outstanding scholarly publications, and his constant use of sophisticated economic thinking bear witness to this. But there is likewise no doubt that Padoa-Schioppa was a particular economist, unique in his own way. Economics was not for him an intellectual exercise in which implications are drawn from axioms through deductive logic. It was a method to analyse problems and to devise solutions that should constantly be tested in real life and discarded, if necessary, in a Popperian sequence. Economics should adapt to circumstances – not the other way round.

Padoa-Schioppa displayed his brilliant economic intuition on many occasions. In his article in *International Finance* entitled 'The crisis in perspective: the cost of being quiet' he described, already in 2008, the salient features of the crisis that surfaced only a year earlier. He placed the bursting of the housing bubble in the USA in an international context, linking it with domestic and foreign imbalances.

His clashes with doctrinaire economists (often classical, but not always) are proverbial. His view on the role of exchange rates as an instrument for adjusting economic imbalances turned out to be central to his career. The doctrine, for which several Nobel Prizes have been awarded, is that exchange rates are, in a wide variety of circumstances, an easier and less painful way to correct accumulated disequilibria among countries than changes (notably, reductions) in nominal wages and prices at national level. Not for Padoa-Schioppa. He viewed exchange rate variability, with the possibility of competitive devaluations and currency wars that it entails, as a fundamentally uncooperative and undesirable way to settle economic divergences, hence his preference for monetary cooperation and ultimately monetary union, in Europe and elsewhere. To be durable, cooperation was to be supported by institutions, notably central banks.

The fifth element of Padoa-Schioppa's thinking and practice consists of his conceptions of central banking and his actions as a central banker. This is the area to which he devoted most of his professional life. He did not start his career as a central banker – he often joked, particularly with those he considered not concrete enough, about his early job as a clothes seller – but in the end he personified the quintessential and all-round central banker as nobody else did, though always in an original way. He viewed central banks as complex and multi-faceted financial institutions pursuing the public good along several interconnected paths. It is the polar opposite of the notion of a single-minded and ring-fenced authority regulating interest rates just to achieve stable prices. Price stability is key, but should be filtered through a broader notion of monetary, financial and economic performance. And it should never be attained – typical of modern macroeconomic models – by simplistic decision rules, by some kind of autopilot. The strategy and transmission of monetary policy cannot abstract itself from the interconnections between the financial system as a whole, from the money markets where short-term assets are bought and sold, or from the payment system, through which money is exchanged for goods and services.

The last, but not least, component of his legacy concerns his profound trust in human beings: Padoa-Schioppa invested heavily in younger collaborators because he firmly believed that only by bringing the people of Europe closer together would there be a better world. At the end of his term as Executive Board member of the European Central Bank in 2005 he gave his colleagues and collaborators a paperweight made of Murano glass, on which he had had engraved the caption of a drawing by Goya which he found in the Prado showing a doddery old man, but one who is eager to keep on learning. The engraving reads, in Spanish, 'Aun aprendo', which means 'I am still learning' – reflecting his approach to life.

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

Selected Works

1987. *Efficiency, stability and equity – A strategy for the evolution of economic system of the European community: A report*. Oxford: Oxford University Press.

1994. *The road to monetary union in Europe – The emperor, the kings, and the genies*. Oxford: Oxford University Press.

1995. Economic federalism and the European Union. In *Rethinking federalism: Citizens, markets, and governments in a changing world*, ed. K. Knop. Vancouver: UBC Press.

1999a. *EMU and banking supervision*. Lecture at the London School of Economics, Financial Markets Group on 24 February. Available at http://www.ecb.europa.eu/press/key/date/1999/html/sp990224.en.html

1999b. *Reflections on the globalisation and europeanisation of the economy*. Lecture at the University of Göttingen, Center for Globalization and Europeanization of the Economy, Göttingen, 30 June. Available at http://www.ecb.europa.eu/press/key/date/1999/html/sp990630.en.html

1999c. *Lectio Doctoralis for the attainment of the Laurea Honoris Causa in International Economics of Trade and Currency Markets*. Trieste, 19 November.

2000. *The eurosystem and financial stability*. Speech at the Belgian Financial Forum, Brussels, 10 February. Available at http://www.ecb.europa.eu/press/key/date/2000/html/sp000210.en.html

2001. *Increased capital mobility: A challenge for the regulation of financial markets*. In *The world's new financial landscape*, ed. H. Siebert. Heidelberg: Springer.

2004a. *The euro and its central bank – Getting united after the union*. Cambridge, MA: MIT Press.

2004b. *Regulating finance: Balancing freedom and risk*. Oxford: Oxford University Press.

2004c. *Europe, a civil power*. Federal Trust for Education & Research.

2010a. *The ghost of Bancor: The economic crisis and global monetary disorder*. Louvain-la-Neuve, 25 February.

2010b. *Markets and government before, during and after the 2007–20xx crisis*, The Per Jacobsson Lecture. Basel.

Palander, Tord Folkeson (1902–1972)

T. Puu

Keywords

Bonds; Choice under uncertainty; Duopoly; Factor substitution; Land use; Location theory; Monopolistic competition; Palander, T. F.; Rent; Spatial monopoly; Stockholm school

JEL Classifications

B31

Palander became an engineer at the Royal College of Technology in Stockholm, before studying economics at Stockholm University. He published his dissertation *Beiträge zur Standortstheorie* (Contributions to Location Theory) in 1935. He was appointed Professor of Economics at the Business School of Gothenburg in 1941 and at Uppsala University in 1947. The dissertation is a standard reference in location theory; it was never published in English, though a Japanese edition was published in 1984.

Chapters III–X of the *Beiträge* contain a very detailed discussion of spatial economics from the classics to recent developments up to the date of Palander's own contribution. The Ricardian and von Thünen land rent theories, the Launhardt-Weber theories of location and market area formation are penetrated, and the spatial facets of Hotelling's and Chamberlin-Robinson's then fresh monopolistic competition theories are for the first time given due regard. Palander aims at an integration of classical location and land use theories with modern developments in general economics. Bringing in profit maximization under price-dependent demand and the possibility of spatial monopoly is one step in this direction. Another is the stress on the importance of factor substitution. Palander is extremely critical of Andreas Predöhl's attempts in this direction. His own planned contribution was withdrawn from the manuscript just before its publication, and the problem first got a satisfactory solution with the contribution by Leon Moses in 1958.

Palander's most original contributions are contained in Chapters XII–XIV. The central theme is to extend the classical models, where transportation is assumed to be along straight lines, to more realistic situations where transport rates are distance dependent, or traffic crosses different media, like land and sea. In the last context, the neat refraction law of traffic was discovered. These contributions set the stage for Martin Beckmann's (1952) continuous model of transportation.

One of the most attractive features of Palander's work is the artwork, which must be considered unsurpassed until the advent of computer graphics. Palander brings graphical analysis combined with simple algebra and analysis to perfection.

Palander remained in the USA as a Rockefeller Fellow during 1936, studying Chamberlin's monopolistic competition theory. His own contributions are a brief abstract in English of a presentation at the Cowles Commission conference at Colorado College 1936 (‘Instability in Competition between Two Sellers’) and an article in Swedish, ‘Konkurrens och marknadsjämvikt vid duopol och oligopol’ (‘Competition and Market Equilibrium in Duopoly and Oligopoly’), published in *Ekonomisk Tidskrift* (later *Scandinavian Journal of Economics*) in 1939. Palander’s particular interest was the stability of adjustment processes in classical Cournot and similar types of duopoly.

Palander belonged to the informal group of economists called the ‘Stockholm School’ and wrote an extensive critical review of their methods, ‘Stockholmsskolans begrepp och metoder’ (1941). Palander’s remarks concern in particular the lack of rigorous dynamic analysis.

Palander took a great interest in Keynesian macroeconomics. He edited a translation of the *General Theory* into Swedish in 1945 (*sysselsättningsproblemet*), with commentary, and wrote an extensive mathematical and graphical analysis of the work in 1942. This article, in comparison to the similar works by Hicks and Klein, contains a thorough analysis of all the different variants with the relations expressed in monetary, real and wage units.

Palander’s later work was mainly pedagogical, and his last research interest concerned monetary theory in connection with choice under uncertainty. He wrote a monograph (in Swedish) on the effects of index bonds upon an inflationary economy in 1957. Among his consultancies the most important was to the Swedish Railway Board on fare tariff policy.

See Also

► Location Theory

Selected Works

1935. *Beiträge zur Standortstheorie.* Uppsala: Almqvist & Wiksell.

1936. Instability in competition between two sellers. In Cowles Commission, Research Conference on Economics and Statistics, Colorado Springs: Colorado College. Chicago: University of Chicago Press.

1939. Koncurrens och Marknadsjämvikt vid duopol och oligopol. *Ekonomisk Tidskrift* 41, Pt. *i,* 41(2), 123–45; Pt. II, 41(3), 222–50.

1941. Stockholmsskolans begrepp och metoder. *Ekonomisk Tidskrift* 43(1), 88–143. Trans. as: On the concepts and methods of the ‘Stockholm School’, *International Economics Papers* No. 3. London and New York: Macmillan, 1953.

1942. Keynes allmänna teori och dess tillämpning inom ränte-, multiplikator-och pristeorien. *Ekonomisk Tidskrift* 44(4), 233–72.

Bibliography

Beckmann, M. 1952. A continuous model of transportation. *Econometrica* 20: 643–660.

Moses, L. 1958. Location and the theory of production. *Quarterly Journal of Economics* 72: 259–272.

Paley, William (1743–1805)

A. M. C. Waterman

Abstract

The whole of Paley’s contribution to economics is contained in a single chapter of *Moral and Political Philosophy* (1785). The object of ‘rational politics’ is to maximize ‘happiness’, and Paley argued that this is achieved by maximizing population. Population is determined by the total supply of ‘provisions’ produced by the agricultural sector. The demand (and hence supply) for ‘provisions’ and for ‘luxuries’ are reciprocally determined. As in Mandeville, the taste for ‘ luxuries’ stimulates production. But it also acts in the opposite direction because it deters population. Paley explicitly recognised the optimization problem and was the first economist to do so.

Keywords
Cambridge; Fluxions; Keynes; Mandeville; Paley; Population

JEL Classifications
B31

William Paley was one of the most powerful influences upon intellectual life in Britain and America from 1785 until the late 1850s. J. M. Keynes (1972, p. 79 n. 2) judged that 'Perhaps, in a sense, he [Paley] was the first of the Cambridge economists.' Paley's demand-led analysis of the determination of total output was in some respects even more 'Keynesian' than Malthus's heterodox macroeconomics of 1820; and it may have been this that earned Keynes's approval in 1933 when he was beginning to excogitate his *General Theory of Employment, Interest and Money* (Keynes 1936).

Paley's Life and Work

William Paley was a Yorkshire man, and was mildly derided at Cambridge for speaking Latin with a Yorkshire accent. He was born in Peterborough in July 1743. His father, the Revd William Paley, was then a Minor Canon of Peterborough, but was appointed Master of Giggleswick School in 1745, whereupon the family returned to Yorkshire. Like his father, the younger William went up to Christ's College, Cambridge, matriculating in 1759 and graduating BA in 1763 as senior wrangler [= highest performance among all Tripos candidates, who by that time were required to attempt a written examination in mathematics in addition to the five Latin disputations]. He was elected fellow of Christ's in 1766 and ordered deacon. In 1767 he was ordained priest and graduated MA. He was awarded the degree of Doctor of Divinity in 1795.

For ten years after election as fellow Paley occupied various college offices and played a large part in teaching undergraduates. At that time this was almost entirely conducted in colleges by college lecturers and tutors. Attendance at college classes was compulsory. All undergraduates faced a common curriculum designed to prepare the next generation of clergymen, magistrates and legislators for their public duties in a Christian society: biblical languages and literature, the Latin and Greek classics, and some reading in 'moral and political philosophy'. The small minority of ambitious students who sought an honours degree supplemented these studies with Newtonian 'natural philosophy' and mathematics. Paley was soon known throughout Cambridge as a superb teacher, and many students came from other colleges to attend his lectures. A later commentator wrote of Paley's 'utter inability to be obscure' (Annan 1984, p. 244). In all probability he taught the entire curriculum, with the possible exception of the classics, in addition to mathematics and natural philosophy for honours candidates.

In 1776 Paley was preferred to the rectory of Great Musgrave, Westmorland, and thus was at last able to marry, resigning his fellowship as was then required. His wife, Jane Hewitt of Carlisle, bore him 10 children, two of whom died in infancy. He remained in the diocese of Carlisle for the rest of his career, where his energy and efficiency soon led to promotion as Archdeacon (1782) and Chancellor (1785). But he also held benefices later in the dioceses of Lincoln and Durham, and in 1796 moved to Bishop Wearmouth in Durham whilst remaining Archdeacon of Carlisle. He was exemplary in parochial and diocesan duties, a leader in the campaign to abolish the slave trade, and active in promoting education of the poor. His first wife having died in 1791, he married Catherine Dobson of Carlisle in 1795. Paley died on 15 May 1805 after a lingering and painful illness, during which he completed his last book. He is buried in Carlisle cathedral.

Soon after leaving Cambridge his friends urged him to write up his college lectures. He began with *Moral and Political Philosophy* (1785) which brought instant fame and fortune. It was almost immediately adopted as a required text for all undergraduates at Cambridge, going through 20 English editions by 1814 (15 in Paley's own lifetime) and 10 American editions by 1821. In the USA it remained ‘the most popular text on moral philosophy from the 1790s to the Civil War’ (Haddow 1939, p. 67). Paley followed this success with two books on the New Testament: *Horce Paulina* (1790) and *Evidences of Christianity* (1794), the second of which led both to his doctorate and to valuable preferment amounting to nearly £2,000 a year. His last book, *Natural Theology* (1802), was a characteristically 18th century attempt to demonstrate the existence and attributes of God from the evidence of His work in nature: in this particular case, biological nature. Sidestepping David Hume’s sceptical critique of teleology (McLean 2003), Paley filled his book with detailed and well-informed examples of biological adaptation as evidence of ‘design’.

Paley said that his books had been written in the reverse order of that in which they should be read. *Natural Theology* established the scientific grounds for belief in God. *Evidences* and *Horce Paulina* showed that Christianity was definitive. *Moral and Political Philosophy* expounded the social and political norms of a Christian society. Thanks to Paley, the Anglophone intelligentsia – unlike their counterparts on the Continent – largely took Christianity for granted down to the 1850s. But the appearance of Charles Darwin’s work on evolution suddenly changed all this. Adaptation was no longer evidence of design, hence Paley’s entire structure fell to the ground. Sales of his books dried up. During the decade of the 1860s ‘Christian dogma fell away from the serious philosophical world of England, or at any rate of Cambridge’ (Keynes 1972, p. 168).

Intellectual Context of Paley’s Economic Analysis

Since there can be no obligation to do that which is unfeasible, moral and political philosophy must entail some positive investigation of the economic and social circumstances to which normative principles apply. Therefore an element of what we now call ‘economic analysis’ is often to be found, implicit or explicit, in all expositions of political philosophy, at least since Plato’s *Republic*.

By the 18th century a great deal of what became ‘political economy’ in the 19th century and ‘economics’ in the 20th had begun to circulate in informed circles in France and Britain, and the outlines of a common body of knowledge can be identified. (1) Agriculture normally affords more food than is necessary to feed those who produce it. (2) The cost of production – of food as of all other commodities – will not normally be incurred unless there is an expectation of an adequate return: ‘effectual demand’ is thus a necessary condition of production. (3) Since manufactured goods need inputs from agriculture (food to sustain manufacturers), an urban manufacturing sector can provide a demand for the agricultural surplus. (4) In the same way, a rural agricultural sector can provide demand for a manufacturing surplus, hence the two sectors are mutually sustaining. (5) Labour needed in production is produced by human beings supplied with food (and manufactured necessities). A certain average per capita income of food and other necessities will keep population and workforce constant. At a higher income these will grow and vice versa.

Though it is evident that Paley was familiar with these ideas, and indeed made them the focus of his own analysis, he gave us no help in discovering his sources: ‘I have scarcely ever referred to any other book, or mentioned the name of the author whose thoughts, and sometimes, possibly, whose very expressions I have adopted’ (Paley 1825, p. xiv). In addition to the common core associated in particular with Mandeville, Cantillon, Quesnay and Adam Smith, Paley is seemingly aware of many other elements of 18th century economic thought to be found in Locke, Hume, Berkeley, Steuart and Josiah Tucker. Yet only Berkeley’s ‘walls of brass, fifty cubits high’ appear with attribution (as they do in Malthus). What Paley seems not to know about, or at any rate not to think important enough to teach his under graduates, are (a) price theory as found in Adam Smith, (b) general equilibrium in competitive markets as pioneered by Boisguilbert, and (c) the virtues of *laissez-faire* as taught by the Physiocrats. The last is in marked contrast to his clerical predecessor, the Revd Josiah Tucker (1713–99), whose praise of the self-regulating market economy was later echoed by Smith.

Although Paley's lectures were prepared before the appearance of *Wealth of Nations* in 1776 (the year he left Cambridge), the occurrence of certain passages in *Principles* which read like summaries of Smith's work suggest the possibility that Paley did read it sometime between 1776 and 1785. Paley's remark that population may double in twenty years is found in Smith (1976, p. 479). His account of money, property and power (1785, p. 604) could be a digest of a similar argument in Smith. Adam Smith's famous trio, 'the butcher, baker, brewer' crops up in Paley, as does the assumption that 'the only spring which keeps human labour in motion' is 'the exclusive right to the produce' (Paley 1825, pp. 493, 489). However, any conclusion on the basis of such evidence can only be conjectural.

Paley's Economics

A single chapter of *Principles,* 'Of Population and Provision; and of Agriculture and Commerce as subservient thereto' (Paley 1825 part VI, chap. XI) contains the whole of Paley's economic writing. The most original features are its explicitly utilitarian basis and its careful analysis of optimality.

Paley's *Principles* was virtually simultaneous with Bentham's *Morals and Legislation* (1789). The two were the first to popularise utilitarian arguments in Britain and the latter is sometimes characterized as 'Paley with God left out'.

Paley's 'economics' chapter begins: 'The final view of all rational politics is the greatest quantity of happiness in a given tract of country' (Paley 1825, p. 477). Since 'communities' etc. are mere abstractions 'nothing really exists or feels but individuals'. In any society some will be happier than others, but Paley assumed that the happiness of individuals is additive and may be aggregated, and that the range and distribution of happiness is independent of scale. Hence 'twice the number of inhabitants will produce double the quantity of happiness' (Paley 1825, p. 478). It follows that the happiness of a community or district is proportional to its population, hence *that the object of 'rational politics' is to maximize population.*

Paley assumed that the human food requirement is biologically given and that human populations expand to the limit set by food supply. Hence the policy goal becomes that of maximizing 'provisions'. The production of 'provisions' is determined by aggregate demand: the sum of demand for food by food-producers themselves, by personal servants employed by their landlords, and by the producers of other goods which Paley labelled 'luxuries'. Similarly, the production of 'luxuries' is determined by the sum of the demand for luxury goods by luxury goods producers and by the producers of food. Social arrangements exist for the 'exchange' of food and luxuries. In a sophisticated version of Mandeville's doctrine, Paley explained that luxury stimulates employment and industry: 'The watchmaker, while he polishes the case, or files the wheels of his machine, is contributing to the production of corn as effectually, though not so directly, as if he handled the spade or held the plough' (Paley 1825, p. 496).

Outputs of 'provisions' and 'luxuries' are determined at equilibrium by the reciprocal demands of each sector for the other's product. Population (and therefore happiness) is determined at equilibrium by food production. The policy problem is therefore to identify the factors which determine equilibrium, and to operate where possible on those which increase production of 'provisions'. Paley abstracts completely from all supply constraints: his is a purely demand-driven model, which may explain why Keynes found it so congenial (Waterman 1996, pp. 681–3).

Given production techniques and the biological food requirement, the determinants of equilibrium are the size of landlords' rents and the taste of the whole population for 'luxuries'.

Since Paley is unaware of, or abstracts from, diminishing returns, he has no theory of rent. But he argued strongly for private property in land and for social arrangements that would give landlords an incentive to maximise rent (Paley 1825, pp. 489–90, 516–19). Landlords are assumed to spend all their rents on personal services, thus supporting an unproductive population and contributing to the demand for ‘provisions’.

The taste for luxuries is an important determinant of aggregate demand, but here Paley takes a large step beyond Mandeville. For it is also the case that a taste for luxuries may operate in the opposite direction to *reduce* population. In a passage that is almost certainly the chief source of Malthus’s concept of the ‘preventive check’, Paley remarks that ‘men will not marry, to *sink* their place or condition in society’ (Paley 1825, p. 485). An increase in luxury makes ‘the usual accommodations of life more expensive’ and raises the cost of ‘the established mode of living’ (Paley 1825, pp. 485–6). Marriage and family formation are deterred, so tending to reduce the population and workforce. For Malthus and his successors, luxury works against population. For Mandeville, luxury stimulates population. Paley is unique in recognizing both these effects.

As befitted a senior wrangler well trained in Newtonian ‘fluxions’ [= differential calculus], Paley also recognized what Malthus (1986, p. 102), himself a wrangler, later identified as ‘the problem of *de maximis* and *de minimis* in fluxions, in which there is always a point where a certain effect is the greatest, while on either side of this point it gradually diminishes’. Thus Paley (1825, p. 486) observed that ‘*luxury*, considered with a view to population, acts by two opposite effects; and it seems probable that there exists a point on the scale, to which luxury may ascend ... beyond which the prejudicial consequences begin to preponderate. Though this ‘arithmetical problem depends on circumstances too numerous . . . to admit of a precise solution’, a formal mathematical reconstruction of Paley’s implicit model can be made and conditions specified for maximization of population (Waterman 1996, pp. 677–80, 685). Paley seems to have been the first analyst in the history of political economy to have recognized clearly that optimization lies at the heart of scientific economic thinking.

Paley’s analysis of the reciprocal demand of ‘provisions’ and ‘luxuries’ was supplemented by consideration of technical progress (‘the abridgment of labour’ by ‘mechanical contrivances’), and of the effect of a ‘continual increase’ in the money supply that seems obviously derived from Hume. The immediate effect of technical progress appears from the reciprocal-demand model to be to diminish ‘provisions’ and population. But Paley argued that ‘some more general and remoter consequences’ may ‘increase the demand for work’ (Paley 1825, p. 514) and hence that employment and output will rise on balance.

See Also

- ▶ English School of Political Economy
- ▶ Historical Economics, British
- ▶ Malthus and Classical Economics
- ▶ Malthus, Thomas Robert (1766–1834)
- ▶ Malthusian Economy
- ▶ Mandeville, Bernard (1670–1733)

Bibliography

Annan, N. 1984. *Leslie Stephen, the godless Victorian*. London: Weidenfeld and Nicholson.

Haddow, A. 1939. *Political science in American colleges and universities, 1636–1900*. New York: Appleton-Century.

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

Keynes, J.M. 1972. *Essays in biography* (1st ed., 1933). vol. X, *The collected writings of John Maynard Keynes*, ed. E. Johnson and D. Moggridge, 29 vols. London: Macmillan.

Malthus, T.R. 1986. Observations on the effects of the Corn Laws (first published 1814). vol. 7, *The works of Thomas Robert Malthus*, ed. E.A. Wrigley and D. Souden, 8 vols. London: Pickering.

McLean, M.R. 2003. Did Paley ignore Hume on the argument from design? In *Faith, reason and economics: Essays in honour of Anthony Waterman*, ed. D. Hum. Winnipeg: St John’s College Press.

Paley, W. 1825. *The principles of moral and political philosophy* (1st ed. 1785). vol. IV, *The works of William Paley, D.D.*, ed. E. Paley, 7 vols. London: Rivington.

Smith, A. 1976. *An inquiry into the nature and causes of the wealth of nations* (1st ed. 1776), 2 vols, ed. R.-H. Campbell and A.S. Skinner. Oxford: Oxford University Press.

Waterman, A.M.C. 1996. Why William Paley was ‘the first of the Cambridge economists’. *Cambridge Journal of Economics* 20: 673–686.

Palgrave, Robert Harry Inglis (1827–1919)

Murray Milgate

Keywords
Palgrave, R. H. I.; Central banking; Palgrave's Dictionary of Political Economy

JEL Classifications
B31

Palgrave was born in London, the third of four male children of Francis Palgrave and Elizabeth Turner. He was named after Robert Harry Inglis – an Old Tory, Member of Parliament, and a friend of Palgrave's father. Quite incidentally, this R.H. Inglis edited some works by the economist Henry Thornton. Palgrave was denied the formal education provided for his two elder brothers, instead entering the banking business of Gurney & Co. (in which his maternal grandfather had been a partner) in Great Yarmouth at the age of 16. Palgrave himself subsequently became a partner in the bank, and married in 1859 a daughter of Mr George Brightwen, who was related to the Gurney family.

Family

Palgrave's father was born Francis Cohen in 1788, the son of Meyer Cohen, member of the London Stock Exchange during most of the years that the Ricardo's were members. Francis Cohen altered his name to Palgrave in 1823 upon marriage to Elizabeth Turner – Palgrave being Elizabeth Turner's mother's maiden name. She, in turn, was the daughter of Dawson Turner, a partner in the English country bank Gurney & Co. in Great Yarmouth. Sometime during the second decade of the 19th century, Francis Cohen renounced the Jewish faith and embraced the Christian religion in the form of the teachings of the Church of England. He was a medievalist of some repute, publishing *The Rise and Progress of the English Commonwealth* in 1832, and *The History of the Anglo-Saxons* in 1837. He was Deputy Keeper of H.M. Public Records, was knighted in 1832, and his literary friends included Macaulay and Henry Hallan (who Palgrave was to quote to moving effect in his editorial preface to the final volume of his *Dictionary*).

His first son, Francis Turner Palgrave (1824–1897), is still widely known today for his famous *Golden Treasury of English Lyrics and Verse*, the first edition of which appeared in 1861. He was educated at Charterhouse and Balliol College, Oxford, going up in 1843, and in 1846 acted as assistant private secretary to Gladstone. Between 1850 and 1855 he directed a government teacher training college near Twickenham. Thereafter he was engaged in the Department of Education in London until his retirement in 1884. In 1885 he was elected into the Professorship of Poetry at Oxford (with the support of Alfred Tennyson, whom he had met in his days near Twickenham).

The second son has William Gifford Palgrave (1826–1888), the least 'typical' of the family. After Charterhouse and Trinity College, Oxford, he moved to India, becoming a lieutenant in the 8th Bombay Regiment. He soon converted to Roman Catholicism, entered a Jesuit mission in Madras, and was ordained a priest of the Order. He remained as a missionary in India until 1853 when he was recalled to Rome. Later in that same year he went as a missionary to Syria. The *Dictionary of National Biography* reports that 'he could and did pass without difficulty for a native of the East', adding that 'the often repeated story that he had officiated as Imaum in mosques is without foundation'. When hostilities between the Druse and the Maronite Christians broke out, the Maronites invited him to become their leader – an invitation which it seems he declined. A massacre of Christians in Damascus in June 1881 precipitated his return to Europe. There he reported to Napoleon III on the Syrian situation. This contact led him into an expedition in 1862–1863 across the Middle East. This was financed by the French government, to whom he was to 'report on the state of the Arab attitude' towards France. Subsequent to this venture he returned to England.

At ‘home’ again, he published his *Narrative of a Year’s Journey Through Central and Eastern Asia* in 1865. This was the most widely read narrative on that region until the accounts of T.E. Lawrence appeared on the scene. He broke with the Jesuits, and became a diplomat in the service of the British government. He was dispatched to Abyssinia, and then went as consul to St. Thomas in the West Indies in 1873. There followed postings in Manila (1876), Bangkok (1879), and Uruguay (1884). He died in Montevideo on 30 September 1888.

The youngest son, Reginald Francis Douce Palgrave (1829–1904), was Clerk of the House of Commons. He was the only sibling to survive to see the publication of the *Dictionary* and Palgrave consistently requested Macmillan to forward to him complimentary copies of the work as it appeared.

Works

It is said that ‘as quite a young boy’ Palgrave received from his father a copy of the *Wealth of Nations*, which he treasured throughout his life. That book seems to exert a power so mysterious that few who take it up and study it seriously have been able to avoid the fate of a career in economics. However, his activities at Gurney & Co. in Great Yarmouth, while immersing him in the daily business of economics, delayed the entry of his name into its literature until 1870, when he received the Statistical Society’s Taylor Prize for an essay on local taxation in Britain and Ireland.

The work by which Palgrave’s name will be perpetuated is, of course, his *Dictionary of Political Economy*, one of the finest achievements of Victorian scholarship. Shortly after the publication of its last appendix he was knighted (1909).

Here, we shall consider only his other writings, all of which dealt with some aspect of banking practice or theory. His publications of 1873, 1874 (a and b) and 1877 typify a kind of statistical analysis of central banking, and their results are largely collated and summed up in *Bank Rate and the Money Market* of 1903. Of this book, Schumpeter commented:

> [It] is a masterpiece of the art of making figures speak . . . it is very difficult to formulate particular results but he who peruses this book page by page suddenly discovers that he understands its subject. (1954, p. 1080)

On matters of policy, he opposed bimetallism, opposed the monetary policy of the government in India, pushed for stability in Bank Rate, and was a supporter of the kind of regulations embodied in Peel’s Act of 1844. However, in a review of Bagehot’s *Lombard Street* (1874b) he formulated clearly the idea that the central bank was effectively an arm of government and thus a vehicle through which governments could effectuate monetary policy. The idea of the Bank of England as an autonomous agency governed only by the legislative provisions of its act of establishment was thereby altered. The new conception which was to take root was of a central bank more familiar nowadays than it was at that time.

In 1877, Palgrave became financial editor at the *Economist,* and on Bagehot’s death took over its editorship. He remained there until 1883. He also edited the *Banking Almanac* until his death, and was briefly editor of the *Banker’s Magazine* to which he contributed regularly after 1880.

Palgrave was also closely involved in the public affairs of the nation. In 1875 he gave evidence before the House of Commons Select Committee on Banks of Issue (George Goschen being the Committee’s economic expert) on behalf of the Country Bankers’ Association, and in 1885 he was a member of the Royal Commission on Depression of Trade and Industry. In the memorandum of evidence submitted to that Commission, Alfred Marshall remarked that he would not cover matters already dealt with in Mr. Palgrave’s memorandum since he was in broad agreement with that document.

It is said that as a boy Palgrave dreamed of becoming a Fellow of the Royal Society – a dream which became reality in 1882, thanks in part to the support he received from Jevons. The latter’s correspondence with Palgrave from that period (held in the archives of King’s College, Cambridge) speaks both to the modesty of Palgrave and the genuine friendship of which Jevons was capable. There is a postcard from Jevons, dated on the day the names of elected Fellows were published in the *Times*, containing no other communication than the name of its adressee: 'R.H. Inglis Palgrave, F.R.S.'

See Also

▶ Palgrave's Dictionary of Political Economy

Selected Works

1870. Local taxation in Great Britain and Ireland. Taylor Prize Essay, Statistical Society of London.

1873. *Notes on banking in Great Britain and Ireland, Sweden, Denmark and Hambourg. London.*

1874a. *Analysis of the transactions of the bank of England for the years 1844–1872. London.*

1874b. Banking. *Fortnightly Review,* 1 January, 92–108.

1877. The influence of a note circulation in the conduct of the banking business. *Journal of the Manchester Statistical Society,* March.

1894, 1896, 1899., ed. *Dictionary of political economy,* 3 vols. London: Macmillan.

1903. *Bank rate and the money market in England, France, Germany, Belgium and Holland: 1844–1900. London: J. Murray.*

Bibliography

Kiddy, A.W. 1919. Obituary – Sir Inglis Palgrave. *Economic Journal* 29: 112–117.

Schumpeter, J.A. 1954. *History of economic analysis*. New York: Oxford University Press.

Palgrave's Dictionary of Political Economy

Murray Milgate

Keywords
Edgeworth, F. Y.; Higgs, H.; Marshall, A.; Palgrave, R. H. I.; *Palgrave's Dictionary of Political Economy*

JEL Classifications
B1

Inglis Palgrave's *Dictionary of Political Economy* appeared in volume form sequentially in 1894, 1896 and 1899. However, 1894 was not the year in which the *Dictionary* began publication. Under an earlier publishing plan, subsequently abandoned, a first part of the *Dictionary* (covering the entries Abatement to Bede) appeared in 1891, followed by two more in the next year (extending the project well into the letter C). Furthermore, 1899 does not accurately represent the completion date of the work. It was not until 1908, when the appendix to the third volume was published, that its publication could be said to have been complete. It took 17 years to effect the publication of the *Dictionary* – better than 20 years of work if one takes into account the fact that the contractual agreement between Palgrave and Macmillan is dated 1888.

Though the original contract called for a work in two volumes, it seems that this plan was subsequently revised to entail publication in parts, each of 120–130 pages in length, and to appear at quarterly intervals. It was envisaged that the entire work would run to between 12 and 14 parts.

The rationale behind the adoption of this plan seems to have derived from a number of considerations. In the first place, the French *Dictionnaire d'èconomie politique* and the German *Handswortsbuch der Staatswissenschaften* had already been appearing in parts, and Palgrave specifically cited these instances in support to the plan. Closer to home, the *Dictionary of National Biography* was also appearing in parts at the time, and successfully at that. Commercial considerations exerted due influence. Palgrave argued that 'each *part* of the Dictionary, as it comes out, may be expected to be noticed . . ., each *volume* would only receive a similar notice', so that '*parts* will be more frequently [brought] before public notice'. He was also concerned that any delay in commencing publication might allow competitors to beat him to the market.

The fact of sequential publication, whether in parts or in volumes, went hand in hand with sequential writing and sequential planning. In 1892, after this process had actually begun, Palgrave argued that it brought two substantial externalities – he might receive 'a good many valuable suggestions and hints', and he could easily refer contributors to the later parts back to earlier published parts in order to avoid overlap. But there were diseconomies as well. With the prospect of publication extending over four or five years, even if it was kept to a strict schedule, there was a real danger of contributor exhaustion. Perhaps more importantly, there was no way of making adjustments for recent advances, or for correcting oversights, or for taking advantage of the valuable advice an editor receives, if the relevant material had already gone to press.

The *original Dictionary* does seem to have suffered some of these disadvantages. Soon after 1899 it was in need of revisions substantial enough to require the printing of separate appendices (published separately at first, and bound in with subsequent reprints). Some of the reasons for this will be touched upon later. What is more, just how exhausted its contributors became during the process can be witnessed in the record of one of the most loyal among their number – F.Y. Edgeworth. In the first volume there were 77 entries from his pen. In the second, the tally had fallen to 38. In the third volume it was down to 17 (in the Higgs edition there are 10 more, mostly addenda to existing entries). Not only did Edgeworth's entries shrink in total number, they shrank in average length as well.

While little concrete detail is readily to hand concerning the editorial practices that Palgrave adopted, there is sufficient evidence available from which to make some fairly confident conjectures. To begin with, it is clear that the list of entries was planned well ahead, despite the more immediate horizons imposed by the choice of sequential publication. In November 1889, for example, before anything had been published, Palgrave reported to Macmillan that he had planned the list of entries down to the letter K. In a letter of 16 March 1892, he reported that he had 'forwarded a considerable number of articles . . . in the S' to the printer. Yet it is equally clear that there was no attempt to generate a list of entries that was in any significant sense 'complete' prior to the commissioning of contributors.

Just how Palgrave arrived at the actual entries to be included is not so easily established. It seems to have been a combination of his own ideas, and those of specialist contributors in particular fields. The list of entries classified by contributor to the original *Dictionary* (compiled by K. Newman and appearing as an appendix of the present work) reveals a pattern whereby certain contributors wrote nearly every entry in a given field. It seems likely that Palgrave simply gave them a free hand to generate the key entries in that field. This probably explains some of the singularities of the pattern of contributions in the original work. How else, for example, might one explain the fact that Mr F.E. Allum of the Royal Mint at Perth, Western Australia, contributed over 100 entries on various media of exchange – from the English Angel to the Japanese Yen. Or that A. Courtois *fils* contributed a similar number of biographical entries on (mainly) French writers – from the marquis d'Audiffret to Louis François Michel Raymond Wolowski. How free were the contributors' hands in determining the length of entries is indicated, perhaps, by the fact that M. Courtois *fils* produced two-and-a-half columns on the obscure Wolowski, and just two on Quesnay.

If the practice of deferring to 'specialists' seems to have introduced certain idiosyncrasies, in other cases it bore fruit, that of Edgeworth being exemplary. It is hardly necessary to add that even with his specialists, Palgrave experienced the usual problems of tardiness in the delivery of sacredly promised entries, and of restraining contributors to limits, even if flexible, as to length. On the whole, however, he seems to have handled these with admirable tolerance and forbearance – though at one point he did suggest that Macmillan might consider sending a man round to Robert Giffen's residence to await on the delivery of his promised essay on Bagehot (which, as it transpired, he did not obtain) – and with not a little creativity in the re-titling of entries so that they would appear further down the alphabet.

The reaction to the *Dictionary* can be considered from two rather different perspectives, that of the economics profession itself and that of the market. As to the latter, there were two phases, the first covering the three parts which appeared before the first 'recognized' volume in 1894, the second covering the subsequent period.

The first phase was a clear commercial failure. The first part had gone to the printers in June of 1891 and was published that year. Within a few months Palgrave was already alert to its lack of success in the market. In a letter to Macmillan dated 14 March 1892 he expresses himself 'extremely disappointed to find that the sale of the Dictionary has been so small'. At about the same time, the publishers began to suggest abandoning the existing publishing plan, in favour of a format more like that which actually appeared. Initially, Palgrave held out against these suggestions. But the similarly disappointing sales of the second and third parts, which appeared in 1892, seems to have reconciled Palgrave to a change of plan. In November of that year little sign remains of the vigorous defence he had made of the earlier plan just a year before. Instead he writes: 'I do not wish *myself* to suggest a change from parts of volumes, but . . . as you appear to have this in your mind, I have now planned out the work as far as the next volume would extend, should *you* desire it be dealt with *as a whole*'.

The professional reaction to the *Dictionary* was generally favourable, as might have been expected given the fact that almost all economists of any repute had already endorsed the enterprise by agreeing to contribute. Of course, any encyclopedia is vulnerable to criticism. Why one particular title for an entry, rather than some other? Why include some unimportant subject or author, and neglect other more worthy ones? There was also some critical comment on the work's sequential appearance. Ever a supporter, Edgeworth effectively put paid to this avenue of attack: 'not even Homer brings forward all his Greeks at once, but makes one the hero of the third, another of the fifth book' (1892, p. 525).

Probably more to the point, some reviewers were wary of the presence in the *Dictionary* of so much material on legal matters, current commercial practices and international treaty arrangements (a hangover from McCulloch's *Commercial Dictionary* perhaps?) and statistical information. This sentiment was shared even by Henry Higgs, who in the editorial preface to his edition of the *Dictionary* remarked that most of this is 'only remotely connected to economics'. The presence of these subjects probably reflected in substantial measure the tastes of Palgrave himself, a commercial banker.

Two specific and less favourable reactions to the *Dictionary* must be singled out, the first contained in an essay by E.R.A. Seligman (who would later edit the *Encyclopedia of the Social Sciences*) that appeared in two parts in the *Economic Journal* for 1903 under the title 'On Some Neglected British Economists'. The second reaction was that of Alfred Marshall.

Seligman's article (in Seligman 1925, pp. 65 ff.) seems to have been an impetus to some of the revisions which Palgrave began shortly after the third volume had appeared. Seligman used the *Dictionary* to exemplify his claim that insufficient attention had been given to a number of British economists: Torrens, W.F. Lloyd, Bailey, Longfield, Read, Craig, Butt and George Ramsay. The cases of Torrens, Lloyd and Longfield were compelling, and Palgrave sought remedy in the appendices. The cases of Craig, Butt and Ramsay were much less so, and Bailey had in fact been the recipient of a generous notice by Edgeworth in the original (where Read had been noticed by James Bonar).

Marshall's reaction, though contained only in asides in correspondence, was unambiguously negative – at one point he makes a play on Palgrave's initials (RIP) in regard to his expectations for the fate of the enterprise. This might help explain why Marshall's name is the most glaring absence from the list of contributors to the *Dictionary*. It was certainly not because he was not asked – it is clear from Palgrave's correspondence with Macmillan that Marshall was approached. In the end the *Dictionary* had to wait for a contribution which bore the signature of Marshall until the Higgs edition of Volume I in 1925. Even then, it was merely a note on the teaching of economics at

Cambridge, not written originally for the *Dictionary* in any case.

The judgement of the profession on his *Dictionary* is probably best summed up in a letter published in the *Economic Journal* for September 1917, congratulating Palgrave on reaching his 90th birthday. For there even Marshall's name appears among the distinguished list of signatories.

Little need be said here of the Higgs edition of the *Dictionary*, which for the first time formally added Palgrave's name to the title (though Edgeworth had done so informally in his review of its second and third parts in 1892), but which made few changes to its structure or contents. Like its predecessor the Higgs edition also appeared sequentially, though not in alphabetical order. Volume II appeared first, in 1923, Volume I followed in 1925, and Volume III in 1926; just why, is not known.

Palgrave was already in his sixties when he began the *Dictionary*, and was in his eighties when it was done – an act of dedication to the discipline unlikely to be replicated. As Edgeworth presciently remarked when reviewing its second and third parts in the *Economic Journal* for 1892, it 'will remain a monument of what may be accomplished by individual initiative and energy'.

See Also

▶ Palgrave, Robert Harry Inglis (1827–1919)

Bibliography

All quotations from Palgrave's correspondence over the *Dictionary* are taken from material held in the Macmillan archive in the library of the University of Reading.

Edgeworth, F.Y. 1892. Review of *Dictionary of political economy*. Edited by R.H. Inglis Palgrave, F.R.S. *Economic Journal* 2: 524–525.

Kiddy, A.W. 1919. Obituary: Sir Inglis Palgrave. *Economic Journal* 29: 112–117.

Price, L.L. 1891. Review of *Dictionary of political economy*. Edited by R.H. Inglis Palgrave, F.R.S. *Economic Journal* 1: 605–608.

Seligman, E.R.A. 1925. *Essays in economics*. New York: Macmillan.

Palmer, John Horsley (1779–1858)

Anna J. Schwartz

Keywords

Bank Charter Act (1844); Bank of England; Bank rate; Central banking; Palmer rule; Usury Laws

JEL Classifications

B31

Governor of the Bank of England from 1830 to 1833, Palmer was a significant participant in 19th-century controversies concerning the Bank's proper management. In 1802 he entered into a partnership with two others as East India merchants and shipowners, and remained active in business until weeks before his death in 1858. '[A] vigorous, outspoken man' (Clapham 1944, vol. 2, p. 114), he was first elected a director of the Bank in 1811 and was regularly re-elected thereafter except for the usual hiatus every third year before 1828 and again in 1845–6. By 1857, when his service terminated, he was the senior director of the Court.

Palmer's view over the period from 1832 to 1857 may be gleaned from his answers to questions addressed to him by Parliamentary committees, three pamphlets he published in 1836 and 1837, and correspondence. Among the central issues he discussed were the nature of the Bank's responsibilities, its relation to the London money market, the joint stock and country banks, its role in stabilizing domestic economic conditions, and how it operated on the exchanges.

Palmer's initial statement of the principles that guided the Bank's policy became known as the Palmer rule or the rule of 1832: the Bank's duty in ordinary times, when the reserve was at a maximum and exchange rates were at par, was to maintain its bullion reserve at one-third of its liabilities, the sum of deposits and note issues. At such times the Bank should hold interest-earning assets of government stock and other long-dated securities equal to two-thirds of its liabilities. Thereafter the portfolio should be maintained unchanged so that as gold was withdrawn from or brought to the Bank, the public would reduce or increase notes and deposits. Changes in the Bank's liabilities would thus arise at the public's, not the Bank's initiative. A loss of bullion would be matched by a reduction in notes and deposits with no change required in the portfolio to reduce the supply of funds in the money market. Palmer held that the Bank should set its rate above the market rate so that in normal times it would not be competing for discounts with London bankers. 'At times of discredit', however, when the market rate rose to the level of Bank rate, the Bank should discount bills of exchange, selling off government securities as discounts increased. (Palmer did not recognize that selling securities would offset the provision of funds to the discount market.)

The validity of the Palmer rule was challenged by both contemporary and later critics (Loyd 1837; Viner 1937), although some modern students (Horsefield 1949; Matthews 1954) regard it as essentially sound. A different line of criticism, occasioned by financial market stringency in 1836–7 and again in 1839, was that in practice the Bank did not observe the rule. Palmer defended the Bank, arguing that in face of deposit declines at other banks associated not with gold drains but with transfers to the Bank – he was referring to East India and other special deposits, 1833–7, and to seasonal Exchequer deposits – it was proper for the Bank to increase its portfolio to offset the extra funds it held. Similarly, the increase in the Bank's portfolio in 1836–7 was the correct response to an internal drain of gold, as it was also in 1839 to an external drain, none of these ordinary years in which the rule applied.

Palmer's arguments failed to convince his interlocutors. He himself retreated from some initial positions. In 1848 he qualified the rule governing reserves in relation to total liabilities. External drains affecting exchange rates, he noted (British Sessional Papers 1847–8, VIII, Pt. 1, 167–8), might be related to political factors abroad rather than to domestic circulation and deposits. As reserves of London bankers gradually came to be held at the Bank rather than as Bank notes, he shifted from a view of Bank rate as the means for influencing note issues to the view that it was the means for influencing the money market. Initially insistent that the Bank must have a monopoly of the note issue – he claimed in 1837 that the issues of many newly formed joint stock banks in 1835–6 had nullified the Bank's contraction of its issues (clearly not true) – in 1848 he did not object to unrestricted country bank note issues provided they were adequately secured.

On other matters, Palmer's views held firm. He believed that changes in Bank rate in relation to rates abroad could control international trade and capital movements. He was a critic of the Bank in the 1840s for too often changing Bank rate and failing to maintain it above market rates. Despite acknowledging the Bank's influence on economic affairs, he denied that it was answerable to anyone but its proprietors, or that publication of a statistical account of its actions was desirable. He opposed separation of the Bank into Issue and Banking Departments before and after the Act of 1844 became law.

Horsley Palmer's name survives as a spokesman for the proper conduct of monetary policy in a period of imperfect understanding of the Bank of England's responsibilities. By asserting that the Bank ought not to compete with other banks in discounting commercial bills in ordinary times, he centred attention on the position of the Bank as distinct from that of other institutions in the money market. He recognized the primacy of its central banking function as lender of last resort during financial stringency. His advocacy of setting Bank rate above market rates hastened the demise of the Usury Laws. For modern observers of the instability that discretionary central bank policies at times have produced, his rule of a constant portfolio has resonance.

See Also

- ► Banking School, Currency School, Free Banking School

Selected Works

1836. *Reasons against the proposed Indian Joint Stock Bank, in a letter to G.G. de H. Larpent Esq.* London: Pelham Richardson.
1837a. *The causes and consequences of the pressure upon the money-market.* London: Pelham Richardson.
1837b. *Reply to the reflections of Mr. Samuel Jones Loyd suggested by a perusal of Mr. J. Horsley Palmer's Pamphlet entitled 'causes and consequences of the pressure upon the money-market'.* London: Pelham Richardson.

Bibliography

British Sessional Papers, session year, Parliamentary Paper number, volume number, pages of Palmer's testimony: 1831–2 (722), VI, 7–70. 1840 (602), IV, 103–41. 1847–8 (395), VIII, Pt. 1, 147–68; Pt. 2, 521–3; 1847–8 (565), VIII, Pt. 3, 79–124. 1857, 2d sess. (220), X, Pts. 1, 2, App., 6 (Palmer letter, dated 15 October 1856); Index, 452.
Clapham, J.H. 1944. *The Bank of England: A history, 1694–1914*. Cambridge: Cambridge University Press.
Horsefield, J.K. 1949. The opinions of Horsley Palmer. *Economica* 16(62): 143–58.
Loyd, S.J. 1837. *Reflections suggested by a Perusal of Mr. J. Horsley Palmer's Pamphlet*. London: Pelham Richardson.
Matthews, R.C.O. 1954. *A study in trade-cycle history: Economic fluctuations in Great Britain, 1833–1842*. Cambridge: Cambridge University Press.
Viner, J. 1937. *Studies in the theory of international trade*. New York: Harper.

Panic of 1907

Jon R. Moen and Ellis Tallman

Keywords
Bank Panic of 1907; Central banking; Federal Reserve Act; National Banking Era

JEL Classifications
N3

The Bank Panic of 1907 was the final banking crisis of the National Banking Era (1863–1913); it was significant in that it led to the Federal Reserve Act. The panic began when the spectacular attempt by F. Augustus Heinze to corner the stock of United Copper Company collapsed on 16 October 1907. The collapse revealed the extensive links of Heinze to another notorious financier in the New York City banking community, Charles F. Morse, a man O. M. W. Sprague (1910, p. 248) describes as having 'an extreme character, even by American speculative standards'. Solvency concerns led to a series of bank runs at several national banks controlled by the two men. Yet the turmoil surrounding the Heinze collapse did not produce a systemic panic in New York, because the New York Clearinghouse took prompt corrective actions on the member institutions.

But on Monday 21 October the National Bank of Commerce announced late in the afternoon that it would no longer clear checks through the New York Clearinghouse for the Knickerbocker Trust Company. The following day, Knickerbocker Trust faced a run on deposits that lasted three hours, and it suspended business just before noon after having paid out $8 million in cash. The next day, the *New York Times* reported that the Trust Company of America was the current 'sore point' in the panic, a report that magnified the run on the Trust Company of America. Over the next two weeks the Trust Company of America paid out $47 million to depositors.

J. P. Morgan, James Stillman of National City Bank, and George Baker of First National Bank arranged through the New York Clearinghouse to aid the Trust Company of America, other stricken trusts, and the stock market after it had been determined that Knickerbocker could not be saved. On 26 October the New York Clearinghouse authorized the issuance of clearinghouse loan certificates to make available otherwise illiquid reserves and assets as a substitute for currency in transactions between banks, currency that could then be paid out to depositors. In addition, the Clearinghouse authorized restrictions on the payments of cash from deposit accounts, thereby limiting the outflow of the means for payment finality from the intermediation system.

While private sector efforts succeeded in quelling the panic, it still altered how key New York bankers perceived their ability to manage the New York money market. New risks set 1907 apart from the earlier national banking era panics of 1893, 1890, 1884, and 1873. These new risks arose from the increased participation of trust companies and other intermediaries in the call money loan market, the overnight loan market for the New York Stock Exchange. By 1907 the New York trust companies had nearly 90 per cent of the loan volume of the New York national banks (Moen and Tallman 1992). As intermediaries outside the Clearinghouse such as foreign banks, national banks from the interior of the United States, and the New York City trust companies became more prominent in the call loan market, the New York bankers realized that their ability to manage that market during panics had diminished considerably.

Under normal financial conditions the call loan market had been reasonably stable, serving as an outlet for excess reserves accumulating in New York through the correspondent banking system. During bank panics, however, the call loan market magnified the scramble for cash as numerous banks tried to call in their loans simultaneously. Typically, interest rates would spike upward and the value of the stock collateral would start to fall as borrowers as a group sold off their collateral to pay call loans. Until 1907, however, such a catastrophe had been averted because the New York national banks, acting through the Clearinghouse, jointly responded to mitigate disruptive contractions of call loans. The collective action of New York national bankers through the issuance of clearinghouse loan certificates and the partial suspension of convertibility of deposits into cash succeeded when clearinghouse banks provided the largest source of funds for the call loan market. The New York banks had been the main suppliers of call loan money since the inception of the stock market, which motivated them to preserve both the stock and call loan markets.

Outside intermediaries like the trusts, having no collective concern for the call loan market, began calling in large numbers of loans on 24 October 1907 (van Cleveland and Huertas 1985). Stock equity values began falling precipitously. Depressed stock values threatened the financial condition of both borrowers and lenders of call loans, including the national banks. The positive correlation between changes in collateral values and the changes in the creditworthiness of borrowers and lenders (sometimes referred to as covariance risk) transmitted the financial shock faced by trusts throughout the financial system to the national banks. Early in the panic there were reports of New York Clearinghouse member banks taking over a large volume of call loans made by trust companies to prevent the collapse of the call loan market and, hence, the stock market. By the end of the panic, loans (and deposits) at the New York national banks increased by nearly ten per cent, in contrast to the 37 per cent contraction of loans at trusts. No similar pattern was seen in the earlier panics (Moen and Tallman 1992).

Contemporary observers applauded the leadership of New York Clearinghouse banks. Sprague (1910) noted how the national banks forbore on calling in loans that were technically insolvent, expecting them to recover as conditions returned to normal. Woodlock (1908) noted the increasing and destabilizing role of outside lenders in the call loan market, lenders like the trust companies that were outside the influence of the Clearinghouse. New York Clearinghouse bankers lost confidence that they alone could reliably alleviate stress in the call loan market during a panic. The movement to establish a centralized system of reserves or a central bank gained momentum after the Panic of 1907 because the New York Clearinghouse banks were no longer willing to bear all the risks alone.

See Also

- ▶ Banking Crises
- ▶ Economic History
- ▶ Federal Reserve System

Bibliography

van Cleveland, H.B., and T. Huertas. 1985. *Citi bank 1812–1970*. Cambridge, MA: Harvard University Press.

Moen, J., and E. Tallman. 1992. The Bank Panic of 1907: The role of the trust companies. *Journal of Economic History* 52: 611–630.

Sprague, O. 1910. *History of crises under the National Banking Era.* National Monetary Commission. Washington, DC: Government Printing Office.

Woodlock, T. 1908. The stock exchange and the money market. Reprinted from *The Currency Problem and the Present Financial Situation.* New York: Columbia University Press.

Pantaleoni, Maffeo (1857–1924)

G. Becattini

Keywords

Marginalism; Pantaleoni, M.; Pareto, V.; Public finance; Regional economics; Value theory

JEL Classifications

B31

Italian economist and politician, Pantaleoni was born in Frascati (Papal States) on 2 July 1857 and died in Milan on 2 October 1924. His career as a university teacher in economics was rather stormy on account of his impatient rejection of any attempt to interfere with his teaching and the free expression of his thought. Elected in 1900 as a radical to the Italian Parliament, he resigned shortly afterwards.

In 1920 he was appointed to manage the finances of d'Annunzio's Free State of Fiume, and in 1923 was nominated a member of the Italian Senate by the Fascist government, of which he was a supporter. His contribution to scholarship may be divided into three parts. First, a famous textbook, *Principii di economia pura* (1889), which contributed to the introduction of marginalist ideas into Italian economic thought and which, in its English translation (1898), made a considerable impression outside Italy as well. Second, a monograph on applied economics on the fall of the 'Credito Mobigliare', which Piero Sraffa aptly compared to Bagehot's *Lombard Street*. And third, a long series of papers, some of which may be regarded as seminal, on a wide variety of topics in both economics and at the interface with other social sciences. His lectures at the University of Rome, transcribed and published by his students, are also worthy of mention.

Much of Pantaleoni's writing has been brought together in anthologies and Pantaleoni's thought has been the subject of much comment; however, no persuasive, thorough, study has yet appeared. The man and the scholar emerge most vividly from his correspondence with Vilfredo Pareto.

The most distinctive feature of Pantaleoni's theoretical work is his tendency to generalize across disciplinary boundaries. Economics, sociology, anthropology and psychology form a kind of unified field within which Pantaleoni, while still employing the style of reasoning of the economist, moves freely and creatively, without shrinking from paradox and logical extremes. His great friend Vilfredo Pareto reproached him in 1898: 'the advancement of scholarship lies in creating new distinctions and not, as you seek to do, in reducing their number.' In apparent contradiction to this extreme tendency towards generalization is Pantaleoni's capacity for minute and piercing analyses and broad and brilliant syntheses of given concrete situations.

How far Pantaleoni may be classified as a genuine marginalist is still an open question. Musgrave and Peacock (1967) wrote that 'one of the first attempts at dealing with the determination of the tax-expenditure plan as a problem of economic value appears in Pantaleoni's essay of 1883'. They refer to the early Pantaleoni paper, 'Contributo alla teoria del riparto delle spese pubbliche', later republished in the *Scritti vari* anthology.

It is true that his extreme subjectivism brings him close to the 'classical' marginalists (though he was very critical of Menger at times), but his eclecticism – in a half-Marshallian vein – about the theory of value, and his acceptance of many of the concepts typical of evolutionist sociology (for example, his distinction between predatory, parasitical and mutualistic settlements), incline one to define him as unclassifiable except in the historical context. His relationship to the thought of Edgeworth and Marshall comes out clearly from the following letter to Edgeworth: ‘you are the closest approximation of a match for Marshall living in England. You know that to my mind, Marshall is simply a new Ricardo who has appeared in the field.’

If we look at Pantaleoni’s mature work, we can conclude, with G. di Nardi, that ‘the Pantaleonian essays following *i Principii,* place him outside orthodox marginalism and made of him a very acute forerunner of contemporary critical schools’.

Pantaleoni’s many disciples have helped to consolidate the profound imprint left by him (much deeper than that of Pareto, though the latter is better known nowadays outside Italy) upon Italian economic thought, especially upon general economic theory and the theory of public finance. Pantaleoni can also be considered among the founders of the modern Italian statistical school and a true forerunner of regional economics. A good example of the most typically Pantaleonian style of reasoning is given by his analysis of the concepts of ‘strong’ and ‘weak’ in economics, so alive and thought-provoking a century later.

See Also

- ▶ Marshall, Alfred (1842–1924)
- ▶ Pareto, Vilfredo (1848–1923)

Selected Works

1889. *Principii di economia pura.* Florence: G. Barbèra. 2nd ed., 1894. Trans. T.B. Bruce as *Pure economics*. London: Macmillan, 1898.

1890. Letter to Edgeworth on 15 November 1890. In the possession of David E. Butler, Nuffield College, Oxford.

1898. An attempt to analyse the concepts of ‘strong’ and ‘weak’ in their economic connection. *Economic Journal* 8: 183–205.

1904–1910. *Scritti vari di economia*, 3 vols, Vols. 1 and 2, Milan: R. Sandron; vol. 3, Rome: Libereria Castellani Editrice.

1925. *Erotemi di economia*, 2 vols. Bari: G. Laterza.

Bibliography

Di Nardi, G. 1977. Pantaleoni rivisitato. In *Annali della Facoltà’ di Giurisprudenza.* Facoltà’ di Macerata. Repr. in G. Di Nardi, *L’Economia a una svolta difficile.* Milan: Giuffré, 1984.

Loria, A., and P. Sraffa. 1924. Maffeo Pantaleoni: Obituary. *Economic Journal* 34: 653–654.

Musgrave, R.A. and Peacock, A.T., eds. 1967. *Classics in the theory of public finance.* London/New York: Macmillan/St Martin’s Press.

Pareto, V. 1960. *Lettere a Maffeo Pantaleoni,* ed. G. De Rosa, 3 vols. Rome: Storia ed Economia.

Papi, Giuseppe Ugo (Born 1893)

F. Caffé

Born in Capua on 19 February 1893, Papi graduated in law in 1913 and pursued an academic career. From 1927 onwards he taught economics as Professor of Political Economy first at the universities of Messina, Palermo and Pavia, and finally in the Faculty of Law at the University of Rome. During his career he was General Secretary of the International Institute of Agriculture from 1938 to 1948, and, subsequently, General Secretary of the Italian committee of the Food and Agriculture Organization. These interests were reflected in the monographs Papi wrote, especially those on the vital importance of agriculture in the process of economic development, and on the emphasis given to free international exchange.

His main contributions were to the theory and politics of the economic cycle, international economics, the theory of the economic behaviour of the state, and the problem of international economic cooperation. In all these areas he remained faithful to his belief in the importance of real factors and of the microeconomic roots of economic phenomena, as compared with the monetary explanations of the trade cycle and of macroeconomic trends. It was characteristic of Papi that while he had no extended knowledge of the new Keynesian theories which dominated the major part of his working life, he continued to believe in the importance of factors governing the allocation of scarce resources and productivity. He believed that these forces could bring about an effective growth in real income.

Selected Works

1923a. *Perstiti esteri e commercio internazionale in regime di carta moneta*. Rome: Signoreli.

1923b. *Il lavatore alla gestione dell'impresa*. Milan: Vallardi.

1932. *Un fattore fondamentale delle fluttazioni economiche*. Riforma sociale.

1933. *Escape from stagnation: An essay on business fluctuations*. London: King & Son.

1935. *La politica della Banca dei Regolamenti internazionali*. Turin: Utet.

1936. *La crisi come negazione di conoscenza*. Padua: Cedam.

1943. *Equilibrio tra attivita economica e finanziaria: Saggi di teoria*. Milan: Giuffre.

1953. *Statistica e macroeconomia*. Milan: Giuffre.

1956. *Teoria della condotta economica dello stato*. Milan: Giuffre.

1959. *Economia internazionale*. Turin: Utet.

1962. *Principi di economia*. Padua: Cedam.

1967. (ed.). *Dizionario di Economia Politica*. Turin: Utet.

References

De Luca, M. 1964. L'oeuvre scientifique de G.U. Papi. *Cahier de l'ISEA*, May.

Paradigms

Peter Urbach

Keywords
Falsificationism; Kuhn, T. S.; Marginalist economics; Paradigm; Popper, K.; Subjective theory of value

JEL Classifications
B4

The idea of a paradigm, in the sense of a dominating principle governing a whole area of scientific research, was invented by Thomas Kuhn (1962), a historian and philosopher of science. One of his starting points was Karl Popper's view of science, according to which the best scientific theories are falsifiable, viz., such theories could, at least in principle, be refuted by empirical evidence. Popper has argued that theories such as those of Freud and Marx, and the central assumptions of astrology, were unfalsifiable, and that in consequence they were inferior to Newtonian and Einsteinian physics, for example. Kuhn, however, pointed out the already well-known fact that the leading theories of the physical sciences are not straightforwardly falsifiable. On the contrary, a theory such as Newton's typically requires auxiliary assumptions before it can make any empirical predictions. If such predictions turn out to be false, then logic alone does not determine whether the main theory or one or more of the auxiliaries is at fault, and a person is then at liberty to retain the central theory and to reject one or more auxiliary.

Indeed, according to Kuhn, this is roughly what scientists generally do. Kuhn was particularly impressed by the way in which the Copernican and Ptolemaic theories, which provided rival accounts of the planetary system, were each sufficiently flexible to account for any astronomical observations. New observations that did not match expectation could always be reconciled with either the Copernican or the Ptolemaic theories by adjusting the system of epicycles, equants, and so on, upon which the planets were supposed to revolve. Kuhn concluded from this that no objective method could determine which theory is the right one or the better one, and that the transfer of allegiance from the earth-centred Ptolemaic to the sun-centred Copernican hypothesis was not based on the rational method of comparing the relative empirical support enjoyed by the two theories: 'the real appeal of sun-centred astronomy was aesthetic rather than pragmatic' (Kuhn 1957, p. 172).

Kuhn described the Copernican and Ptolemaic programmes as rival paradigms. A paradigm, for Kuhn, is a '[way] of seeing the world and of practising science in it' (Kuhn 1970, p. 4). Thus a paradigm incorporates a set of assumptions, especially assumptions about the fundamental nature of some aspect of the world, such as whether there are atoms or not. These assumptions are accepted unquestioningly by adherents to a paradigm. Scientific research is normally conducted in the context of such a paradigm, but it is not directed towards testing the paradigm; on the contrary, it consists in 'a strenuous and devoted attempt to force nature into the [paradigm's] conceptual boxes' (1970, p. 5).

Paradigm-directed research was dubbed 'normal science' by Kuhn to indicate that most work in science is of this kind. Normal science involves fact-gathering activities such as the determination of specific physical quantities (for example, stellar positions), and experiments designed to check theoretical predictions; it also 'consists of empirical work undertaken to articulate the paradigm theory, resolving some of its residual ambiguities and permitting the solution of problems to which it had previously only drawn attention' (1970, p. 27). The crucial feature of normal science is that it is directed to 'puzzle-solving', and 'not [to] major substantive novelties' (1970, p. 35), and predictive failures are normally blamed on the scientist rather than on the paradigm itself.

In addition to a set of assumptions, a paradigm comprises prescriptions for research; however, the precise character of a paradigm is impossible to state. According to Kuhn, the nature of a paradigm resembles the meanings of words, as Wittgenstein characterized them. According to Wittgenstein, a word like 'game' cannot be defined fully and explicitly, its meaning can only be intuited or grasped. One learns the meaning of such words by exposure to examples or paradigm uses. The same applies to a scientific paradigm, which is taught through exemplary or paradigmatic applications of a theory to a concrete range of phenomena. A similar view was held by Polanyi (1958).

How is one paradigm supplanted by another? Kuhn claimed that this is not done by the careful weighing of evidence for each and the selection of the one with the greater empirical support. He argued that this would not be possible, first of all because of the inarticulable and elusive character of a paradigm and secondly because paradigms are 'incommensurable', in that they are really dealing with quite separate phenomena. Kuhn claimed that the observations made by scientists are never 'pure' or 'theory-free', but are always interpreted in terms of the prevailing paradigm. In Kuhn's view, this means that there can be no data that would facilitate a comparison between two paradigms (Kuhn 1970). Hence, Kuhn argued, a rational comparison of the two paradigms is impossible and 'communication across the revolutionary divide is inevitably partial' (1970, p. 169).

Kuhn likened a change of paradigm, or scientific revolution, to a gestalt-switch, calling it 'a transition between incommensurables' (1970, p. 50). After a prolonged period of repeated failures to resolve anomalies in accordance with a paradigm's internal criteria, a time of 'crisis' sets in. Scientists are then prone to change paradigm all of a sudden, after which they see the world in terms of a new paradigm. Sometimes the 'shape of the new paradigm is foreshadowed' during the period of crisis. More often, though, 'the new paradigm, or a sufficient hint to permit later articulation, emerges all at once, sometimes in the middle of the night, in the mind of a man deeply immersed in crisis' (1970, pp. 89–90).

Some aspects of Kuhn's view of science have been widely accepted, in particular, his claim that theories may be tenaciously upheld in the face of apparently unfavourable evidence and that such theories often generate a long programme of research, in which they are defended and extended to new areas.

Other Kuhnian claims have, however, been contested, especially the idea that the nature of a paradigm is inarticulable and that apparently competing paradigms are incommensurable. Lakatos argued that paradigms, or research programmes, as he called them, could be accurately described. In his view, they consisted of a set of unfalsifiable theories (or a 'hard core'), together with a heuristic, or instructions and hints on how to apply the hard core theories to specific explanatory tasks.

The idea that different paradigms are incommensurable has been criticized in a number of ways. The most telling objection is this: although two theories, such as the Newtonian and Einsteinian theories, describe the world in terms peculiar to each and so, in a sense, are disconnected, when combined with appropriate auxiliary assumptions, they may each make predictions that are comparable. Such predictions are not expressed in a pure observation language, but they may be expressed in terms common to both paradigms, and this, contrary to Kuhn's claim, is all that is needed for a comparison of the explanatory powers of rival paradigms. For example, Newton's theory and Einstein's made different predictions over the rate of precession of the perihelion of the planet Mercury and this may be described in a way that is valid and comprehensible, whichever of these theories one favours.

It has been argued that such comparisons permit the relative merits of paradigms to be *rationally* assessed, and allow one to conclude that one of them is 'objectively' the better. Thus Lakatos spoke of a research programme that consistently leads to successful novel predictions as 'progressive', and those that produce a string of failures as 'degenerating'. Lakatos believed that he could advance beyond Kuhn by exploiting this dichotomy to supply rational criteria for paradigm choice. However, he was unable to justify this claim.

Treating scientific research in terms of paradigms or programmes is certainly a useful historiographical tool of analysis. However, there are often difficulties in determining the boundaries of a paradigm. This difficulty means that one often cannot be sure whether a change of view amongst scientists constitutes a revolution, that is, a replacement of one paradigm by another, or whether it is merely a change within a single, more general paradigm (Toulmin 1970).

Perhaps the most significant development in economics that resembles a paradigm is that of marginalism, in particular, the subjective theory of value and the associated methods of marginal analysis. This became the dominant approach of economists from the 1870s on (see Schumpeter 1954, ch. 7). However, a number of authors have picked out parts of this general approach as constituting separate paradigms, or research programmes. For instance, Latsis has given a detailed account of the hard core and heuristic methods of the neoclassical theory of the firm. And Blaug has outlined the paradigm of 'economic equilibrium via the market mechanism', which he traced to Adam Smith. De Marchi has described the Ohlin–Samuelson approach, in which the relative commodity outputs of different nations are connected to factor endowments, as a paradigm. Within macroeconomics there are competing paradigms, namely that which assumes continuous market clearing based on rational expectations (new classical macroeconomics), and that which is based on the old and neo-Keynesian approaches (see Klamer 1984).

A disadvantage of Kuhn's and Lakatos's views is that they suggest that the rules of science allow any group of theories to be set up as the basis of a paradigm. This has led to some areas of inquiry, especially those dealing with social phenomena, mistakenly being regarded as scientific, simply because they are dominated by some tenet that is dogmatically upheld against all difficulties. That such conditions are insufficient for a theory to be acceptable as science has been argued by Dorling (1979). Proceeding from the assumption that theories are judged by their probabilities, in the light of the available evidence, Dorling demonstrated the conditions under which a theory may remain very probable, even when the combination of that theory with some auxiliary hypothesis has been refuted. Redhead (1980) showed how this fact could lead to some theories forming the basis of a paradigm, and how, after a number of unsuccessful predictions, confidence in the assumptions of the programme would gradually be eroded. Thus the crucial characteristics of paradigms may be explained and rationalized.

Bibliography

Blaug, M. 1976. Kuhn versus Lakatos or paradigms versus research programmes in the history of economics. In *Method and appraisal in economics*, ed. S.J. Latsis. Cambridge: Cambridge University Press.

De Marchi, N. 1976. Anomaly and the development of economics. In *Method and appraisal in economics*, ed. S.J. Latsis. Cambridge: Cambridge University Press.

Dorling, J. 1979. Bayesian personalism, the methodology of scientific research programmes, and Duhem's problem. *Studies in History and Philosophy of Science* 10: 177–187.

Klamer, A. 1984. *The new classical macro-economics: Conversations with the new classical economists and the opponents*. Brighton: Wheatsheaf Books.

Kuhn, T.S. 1957. *The copernican revolution*. Cambridge, MA: Harvard University Press.

Kuhn, T.S. 1962. *The structure of scientific revolutions*. Chicago: University of Chicago Press . 2nd edn, 1970.

Lakatos, I. 1970. Falsification and the methodology of scientific research programmes. In *Lakatos and musgrave (1970)*.

Lakatos, I., and A. Musgrave, eds. 1970. *Criticism and the growth of knowledge*. Cambridge: Cambridge University Press.

Latsis, S.J. 1976. A research programme in economics. In *Method and appraisal in economics*, ed. S.J. Latsis. Cambridge: Cambridge University Press.

Polanyi, M. 1958. *Personal knowledge*. London: Routledge & Kegan Paul.

Redhead, M. 1980. A Bayesian reconstruction of the methodology of scientific research programmes. *Studies in the History and Philosophy of Science* 11: 341–347.

Schumpeter, J.A. 1954. *History of economic analysis*. London: George Allen & Unwin.

Toulmin, S. 1970. Does the distinction between normal and revolutionary science hold water? In *Lakatos and musgrave (1970)*.

Paradoxes and Anomalies

N. De Marchi

Keywords

Allais paradox; Anomalies; Arrow's theorem; Capital reversal; Falsificationism; Gibson paradox; Giffen goods; Giffen paradox; Kuhn, T. S.; Leontief paradox; Mill, J. S.; Paradox of thrift; Paradoxes and anomalies; Parametric paradox; Popper, K.; Preference reversal paradox; Private vices, public virtues; Social choice; St Petersburg paradox; Transitivity; Voting paradoxes; Water–diamonds paradox

JEL Classifications

B0

Paradox originally meant contrary to accepted opinion. In logic something more precise is usually intended. A paradox is involved, for example, if we are led to a contradiction by sound reasoning. Economists on the whole seem to have stayed closer to the original sense. We can use this fact to claim that there is ground for treating economists' paradoxes simply as puzzling outcomes. There have been rhetorical appeals to 'paradox', as we shall see; but there are also numerous examples of substantive puzzles. They are to be expected as the limits to existing ways of explaining are explored. This usage has the advantage therefore of allowing us to treat paradoxes as a normal aspect of ongoing inquiry, and it shifts the focus of interest in them away from a status as intellectual curiosities to a status as stimulant to further research.

Anomaly, in the physical and biological sciences, is used to refer to an observational irregularity, or an exception. Economists are reluctant to claim a similar implied degree of regularity, or reliability in estimation. There are, however, at all times facts which resist ready incorporation within existing ways of explaining. We may reserve the term anomaly for these, to indicate the empirical origin of the puzzles which they pose. For the rest, however, there seems little reason to distinguish sharply between paradox and anomaly.

An Interpretative Framework

To make much sense of how paradoxes and anomalies relate to change and progress in economics, we need a framework. It seems probable that economists, like other scientists, are more disposed to admit corrections that they can deal with without altering too radically their existing theories. Challenges which threaten precipitate depreciation of their human capital are likely to be resisted. Furthermore, challenges to 'hard core' propositions, to use Lakatosian terminology, will be resisted absolutely. (This has certain implications for the way 'core' change, if it comes, will be experienced, but these need not occupy us here.) Challenges within the 'protective belt' (in the 'demi-cores', as Remenyi (1979) refers to them, or at the subdisciplinary nodes) will be dealt with more flexibly, but we may expect a different response according to whether the challenge is empirical or analytical in *origin*. Theoretical challenges, so long as they are less than 'core'-threatening, actually provide occasions for the display of ingenuity, and are a major vehicle for change. Empirical puzzles, on the other hand, for reasons that are pretty well understood (but will be reviewed below) seem generally to be regarded as less compelling. The accumulation of empirical anomalies which eventually become so numerous as to crush a theory by sheer weight, and to which Thomas Kuhn (1970) points as a major precipitating factor in revolutions in physical science, is not at all familiar in economics. Gross anomalies, of course, such as unemployment in the 1930s or stagflation in the 1970s, may provoke basic rethinking; but these are not our concern here.

If these suggestions broadly capture the situation and behaviour of economists, we would expect to observe little impact of empirical anomalies on preferred ways of thinking, and relative autonomy in theoretical developments. As Lakatos puts it (Lakatos 1970, p. 136): 'if the positive heuristic is clearly spelt out, the difficulties of the programme are mathematical rather than empirical.'

Such expectations merely reflect the rationalistic conception of their discipline that many economists seem to hold. Historically, that mindset derives from John Stuart Mill. Mill, though in principle a radical empiricist (V.R. Smith 1985, pp. 269–77), managed nonetheless to formulate an economic methodology in which there is no uncertainty, merely incompleteness. This accords well with the hypothetico-deductive model of explanation which economists have known since Mill and have found attractive as re-formulated in recent decades by Sir Karl Popper. On this view, science progresses by the making of bold conjectures, boldness meaning that there is much in the world of 'facts out there' that could refute them; and by subjecting these conjectures to factual tests to identify and help eliminate falsehood. Economists' Millian inheritance leads them to put a particular gloss on this: given our inability to conduct controlled experiments, we tend to look for certainty in premises that we 'know' to be true, by reason of introspection or casual observation. Hence we reason downward *from* truth. In this model, factual evidence can only be at odds with theory if our variables are incorrectly measured, or if we have failed to incorporate all those which are relevant to an explanation, or if the empirical model supposedly corresponding to our theory is incorrectly specified. These attitudes infused the early work of econometricians (Morgan 1984, chs 5, 7); and even the sophisticated methodology of the Cowles Commission in the 1940s used economic theory in a peculiarly Millian manner, to provide a priori grounds for rendering the problems of structure and causation operationally tractable.

With these general considerations in mind we turn to paradoxes and anomalies. What follows is not a survey, nor is there space to examine any single instance in detail (though several receive fuller attention elsewhere in this Dictionary). The instances mentioned serve us as illustrative material and are drawn together in this way in the hope of stimulating further exploration. We shall look at three categories: rhetorical paradoxes; 'fact of life' paradoxes, such as the failure of aggregation rules; and the main group, theoretical paradoxes and empirical anomalies. This last we shall split, as far as seems sensible, into challenges to the hard core and positive heuristics of the dominant neo-Walrasian style of analysis or research programme (for which see Weintraub 1985, ch. 7) and challenges within the protective belt.

Rhetorical Paradoxes

Here use is made of terminological fuzziness, or a premise is left unstated, so as to excite puzzlement and interest in the reader. Adam Smith's *diamonds and water paradox* is of the first sort. Neoclassical economists have, on the whole, viewed the puzzle as emanating simply from a confusion of total with marginal utility. An example of the second sort is again provided by Smith. When he avers that 'it is not from the benevolence of the butcher, the brewer, or the baker, that we expect our dinner, but from their regard to their own interest' (Smith 1776, vol. 1, pp. 26–7), the air of paradox is deliberate. It is dispelled when he goes on to relate the proposition to the principle of occupational specialization. Paradox was a favourite literary device in an earlier age. Donald McCloskey (1985) has recently alerted us to many others in the writings of modern economists.

Paradoxes Arising from the Absence or Failure of an Aggregation Condition

Examples here are the *paradox of thrift, Mandeville's paradoxes* about private vices (for example, profligacy) leading to public virtues (jobs) and *Arrow's impossibility theorem.* The first two, like Smith's paradox of self-interested behaviour leading to socially beneficial results or the Austrian view of social outcomes as complexes of individual choices which interact unpredictably, are instances of unintended consequences. Economists have failed to provide convincing reductionist accounts of aggregative behaviour and tend to take unintended consequences as a fact of economic life. Consistently with the dominant commitment to the neo-Walrasian approach, but paradoxically from any other point of view, this does not stop them from employing micro-motives to account for aggregate relations whose entities they cannot explain. (Excellent discussion of these things is to be found in Elster 1978, ch. 5, and in Nelson 1984.) Arrow's theorem, in so far as it is regarded as a generalization of the paradox of voting (as he himself is inclined to view it) creates difficulties at different levels in different problems (Sen 1985); but far from issuing in defeatism or the rejection of economic rationality it has given rise to a whole new sub-discipline, social choice theory.

Challenges to the Neo-Walrasian Hard Core and Positive Heuristics

Here we shall consider examples of both theoretical and empirical origin and note responses within the profession.

Take first the possibility of *capital reversal* or *reswitching*. In considering an array of techniques in a two-factor, two-product model, capital reversal occurs if, as the wage rate rises (interest rate falls), a less capital-intensive technique is chosen. A far-reaching implication is contained in this simple possibility. If there is no strictly monotonic relation between interest and the capital–labour or capital–output ratio, then it is conceivable that a more, then a less, then once again a more capitalintensive technique is the more profitable as the interest rate declines. This undermines the traditional demand curve for 'capital', negative in the interest rate, since relative goods prices may differ as between two interest rates at which the same technique may be equally profitable. There is then no unambiguous way to value 'capital' (Blaug 1985, pp. 523–8). The very possibility seems to render unserviceable the traditional aggregate production function. Neo-Walrasian economists in effect concede all of this yet go on using devices like the factor-price frontier, suggesting something like absolute resistance to challenges to the basics of the dominant research programme.

A curious instance with somewhat similar implications is the *Giffen paradox.* The positively sloped demand curve was thought of by Marshall as an empirical anomaly, and 'discoveries' of such phenomena by early econometricians led to identification and other 'correspondence' problems being defined (Morgan 1984, ch. 6). It has always been doubtful whether there are any actual observations of Giffen goods, and the strong presumption of theorists and theoretically influenced econometricians has been that, in Stigler's words, 'experience and common sense are opposed to the idea of a positively sloped demand curve' (Stigler 1965, p. 384). Thus even the standard price-theoretic rationalization in terms of a negative income effect dominating weak substitution effects (possibly due to strong rivalry between goods) is quite unaffected by the fact that tests normally turn up positive income effects.

This merely confirms the theorist's suspicion. While the case of Giffen goods is not all that significant, the typical theorist's attitude in this instance is interesting because it is wholly in line with what is observed elsewhere: within the programme, theoretical developments are relatively autonomous. Tests of demand theory were reported in 1975, for example, which in the words of the authors 'make possible an unambiguous rejection of the theory of demand' (Christensen et al. 1975, p. 381). The authors, however, did not refer to their own results as puzzling or anomalous, and their frontal assault also went unremarked.

A third example, involving experimental evidence, is the *preference reversal paradox*. Experimental trials conducted in the 1970s and 1980s have caused consternation by consistently implying intransitivity between individuals' direct preference rankings over risky prospects and the respective certainty equivalents they assign to them. Individuals will choose a high probability of low gain over a low probability of high gain while assigning a higher monetary value or certainty price to the second. This evidence appears to undermine all theories of choice which require transitivity (Machina 1983, pp. 76 ff.). As Imre Lakatos points out, however, 'there is no falsification before the emergence of a better theory' (Lakatos 1970, p. 119); and economists who use standard choice theory have remained impassive in the face of this evidence. The literature openly addressing the matter is apt to challenge the experimental design or to argue that intransitivity of certainty equivalents does not imply intransitivity of preferences (Safra and Karni 1984).

Challenges in the Protective Belt

Here we are much more likely to see positive responses to puzzling outcomes since there is more room for theoretical manoeuvre. Well-known examples among challenges of this sort include the *St Petersburg paradox*, the *Allais paradox*, the *Gibson paradox* and the *Leontief paradox*.

The Bernoulli solution to the St Petersburg paradox has been amended by placing bounds on the utility function. The Allais challenge to the independence axiom of expected utility theory has produced some modification in the specification of the subjects' choices – questioning the experimental design is something of a standard defence – but has issued mainly in the development of non-expected utility models (Schoemaker 1982). It is perhaps worth noticing that the Allais results might have been taken as a challenge to choice theory as much and therefore as threatening to the hard core; but turning the issue into one of model choice has deflected the threat into the protective belt. The Leontief paradox was explained initially as very largely due to omitted or improperly specified variables (De Marchi 1976), but more recent work has focused on a full articulation of the theory that would suffice to generate trade-revealed factor abundance (Leamer 1984, pp. 50 ff.). The Gibson paradox – an observed close correlation between long series of prices and bond yields – was given a complex explanation by Keynes in terms of market interest rate stickiness relative to the natural rate (Keynes 1930, vol. 6, pp. 182–3). Keynes thought of his account as rejecting the simple quantity theory account of the matter; but more recently Gibson's observations have simply been absorbed into the infinite time-horizon, intertemporal choice models of macroeconomics embodying a modified quantity theory approach.

What we see at work in all four of these cases is the emergence of a paradox being taken as an occasion for further theoretical refinement. Only one challenge was analytic in origin, but the three instances of empirical anomaly were treated as invitations to amend theory rather than to abandon it, and in this sense they were subsumed under the power of the positive heuristic. The difference between the way empirical anomalies are handled when the hard core is threatened and when it is not does not turn on any weakening of the rationalistic presumption that theory is the true arbiter. It is simply a methodological choice that is made: challenges to the hard core are inadmissible; challenges to theories in the protective belt allow of a more positive reaction.

Methodological Reasons Why Empirical Results Are Less Compelling

Finally, it is worth asking again why empirical tests result which look 'wrong' generally seem to be judged so, rather than to be taken as falsifications of theory. Theory may be modified by empirical challenges in the protective belt, but is unlikely to be rejected.

One problem is the economist's special version of the Duhem–Quine thesis to the effect that one never tests an hypothesis in isolation, but always in combination with a host of background conditions. Because it is often difficult to know the exact translation of theoretical terms in economic theories into empirical counterparts, and because the data often are not quite what the theory requires, tests are mostly joint tests of theory *and* the adequacy of proxies. A major reason, in turn, why the translation is not clear cut is that economic theories tend to be generic (choice theory, for example). It is possible to model such theories in many, many specific ways; but a negative test result for any one such specification has no implications for the generic theory. A third reason is that we do not observe in economics the constants of physical science. Especially in policy contexts it tends to be the case that altering a parameter of interest causes changes in other parameters. This is a particular form of what Klant has called the *parametric paradox* (Klant 1984, ch. 4.9). Here, and in much comparative statics analysis, our 'constants' are algebraic rather than numerical: they actually function as variables. This robs our models of predictive content unless very special restrictions can be devised (such as cross-equation restrictions in macro policy models) and rendered operational.

In sum, paradoxes, regarded as puzzling outcomes, are normal occurrences in the ordinary line of economic inquiry. Where they impinge on those basic commitments that determine what is 'acceptable' in a line of research there is every incentive to dismiss or ignore them. This is true for analytical puzzles (for example, reswitching) as well as for those in the form of anomalous empirical findings (for example, experimental results indicating preference reversal; Giffen goods; and contrary tests of basic demand theory). It is much easier to modify specific theories in the protective belt, though responses to challenges occurring there will still be driven by the positive heuristics of the research programme (as in the case of the Leontief and Allais paradoxes). The presumption even at this level is that theory arbitrates. There are good reasons for this in the generic nature of economic theories, in the specific difficulties of translating theory into an empirical model and in getting appropriate observations, and in the fact that genuine constants have not been found in economics.

Bibliography

Allais, M., and O. Hagen, eds. 1979. *Expected utility hypotheses and the Allais paradox*. Dordrecht: Reidel.

Amihud, Y. 1979. Critical examination of the new foundations of utility. In *Allais and Hagen (1979)*.

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

Blaug, M. 1985. *Economic theory in retrospect*. 4th ed. Cambridge: Cambridge University Press.

Christensen, L.R., D.W. Jorgenson, and L.J. Lau. 1975. Transcendental logarithmic utility functions. *American Economic Review* 65: 367–383.

De Marchi, N. 1976. Anomaly and the development of economics: The case of the Leontief Paradox. In *Methods and appraisal in economics*, ed. S. Latsis. Cambridge: Cambridge University Press.

Elster, J. 1978. *Logic and society*. London: Wiley.

Harcourt, G.C. 1973. *Some Cambridge controversies in the theory of capital*. Cambridge: Cambridge University Press.

Hayek, F.A. 1931. The 'paradox' of saving. *Economica* 11: 125–169.

Keynes, J.M. 1930. A treatise on money. In *The collected writings of John Maynard Keynes*, vol. 5 and 6. London: Macmillan for the Royal Economic Society. 1971.

Klant, J.J. 1984. *The rules of the game. The logical structure of economic theories*. Cambridge: Cambridge University Press.

Kuhn, T.S. 1970. *The structure of scientific revolutions*. Revised ed. Chicago: University of Chicago Press.

Lakatos, I. 1970. Falsification and the methodology of scientific research programmes. In *Criticism and the growth of knowledge*, ed. I. Lakatos and A. Musgrave. Cambridge: Cambridge University Press.

Leamer, E.E. 1984. *Sources of international comparative advantages: Theory and evidence*. Cambridge, MA: MIT Press.

Leontief, W.W. 1953. Domestic production and foreign trade: The American capital position re-examined. *Proceedings of the American Philosophical Society* 97: 332–349.

Machina, M.J. 1983. The economic theory of individual behavior toward risk: Theory, evidence and new directions. Technical Report No. 433, Center for Research on Organizational Efficiency, Stanford University.

McCloskey, D.N. 1985. *The rhetoric of economics*. Madison: University of Wisconsin Press.

Morgan, M.S. 1984. The history of econometric thought. Analysis of the main problems of relating economic theory to data in the first half of the twentieth century. Ph.D. thesis, University of London.

Nelson, A. 1984. Some issues surrounding the reduction of macroeconomics to microeconomics. *Philosophy of Science* 51: 573–594.

Remenyi, J.V. 1979. Core demi-core interaction; toward a general theory of disciplinary and subdisciplinary growth. *History of Political Economy* 11: 30–63.

Safra, Z. and Karni, E. 1984. 'Preference reversal' and the theory of choice under risk. Working papers in economics no. 154, Department of Political Economy, Johns Hopkins University.

Samuelson, P.A. 1977. St Petersburg paradoxes: Defanged, dissected, and historically described. *Journal of Economic Literature* 15: 24–55.

Schoemaker, P.J.H. 1982. The expected utility model: Its variants, purposes, evidence and limitations. *Journal of Economic Literature* 20: 529–563.

Sen, A.K. 1985. Social choice and justice: A review article. *Journal of Economic Literature* 23: 1764–1766.

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

Smith, V.R. 1985. John Stuart Mill's famous distinction between production and distribution. *Economics and Philosophy* 1 (2): 267–284.

Stigler, G.J. 1965. *Essays in the history of economics*. Chicago: University of Chicago Press.

Weintraub, E.R. 1985. *General equilibrium analysis: Studies in appraisal*. Cambridge: Cambridge University Press.

Pareto as an Economist

A. P. Kirman

It is remarkable that as recently as 1968, Allais in his survey of the contributions of Vilfredo Pareto was able to say that: 'His influence on the development of economics as a science was felt only after considerable delay and has largely been confined to Italy and France.'

While the first part of this statement is not open to objection, most modern economic theorists now regard Pareto as one of the founders of 'economic science'. To see how this reevaluation has taken place, we need to examine Pareto's general approach to economics, certain of his specific contributions and his relationship to the work of his contemporaries and his predecessors. What is clear is that while Pareto's reputation has considerably increased in recent years, this reputation is now built on a very limited part of his work. Thus as he has become better known, the extent of his contribution has been less appreciated. Pareto's professional activity as an economist did not start seriously until 1892 although he had published several articles before that date. This late start, together with the general intellectual environment of the period, explains the curiously varied quality of his work. Although he condemned literary economists out of hand and professed to be interested only by a strictly scientific approach to economics, he frequently made normative judgements and indulged in casual empiricism. Yet his work contains pieces of serious and careful analysis, rigorously worked out, which have had a profound and lasting impact on economics. The three major contributions to economics are the *Cours d'économie politique* (1897), the *Manuale d'economia politica* (1906) revised and translated as the *Manuel d'économie politique* (1909) and his article 'Economie mathématique' in *L'Encyclopédie des sciences Mathématiques* (1911). To these, must be added his articles, in particular those collected and published later as *Marxisme et économie pure* (1966) in the *Oeuvres complets* (1964–84) together with the *Trattato di sociologia generale* (1916) which, with *Les systèmes socialistes* (1902–1903) includes a substantial body of economic analysis. Of these contributions, the first, the *Cours*, originally published in two volumes, contains an exposition of economic theory illustrated with numerous empirical facts. The theory is presented in a more precise and refined way than that of his intellectual predecessor Walras and the emphasis is consistently and unequivocally on the interdependence of economic phenomena and the idea of general equilibrium. Nevertheless, the *Cours* does not seem, to modern eyes, well organized, gives the strong impression of having been assembled from course notes, sometimes without a great deal of reflection, and periodically indulges in direct pleading for the 'liberal' cause. It should, however, be remembered that while so much of the material that Pareto discusses is now standard and has been refined by successive generations of economists, much of it was new, recent, or even original for him and it should be judged in context. What is remarkable is that Pareto, although one of, if not the, founder of the school which culminated in the Arrow–Debreu model, did not hesitate to cast his net wide. He included empirical observations and examples of economic phenomena for which he was able to develop little satisfactory theory. Much of the statistical material in the *Cours* had, according to Pantaleoni (1924), been gathered whilst Pareto was a business man and this fact may have some bearing on its presentation.

The *Manuel*, a deeper and more complete book, illustrates the coexistence of philosophical reflection, empirical observations and rigorous analysis in Pareto's work. Of this book, it is the last section, the 'Mathematical Appendix' (in the French edition, pp. 538–671) which has come to be thought of as Pareto's basic contribution to the theory of general equilibrium and to what we now call 'Pareto optimality' but which he referred to as 'The maximum of society's ophelimity'. This appendix was considerably modified and rewritten for the French edition, in large part as a result of Volterra's (1906) comments on the Italian edition. Although the appendix with its formal analysis is the most widely cited part of the *Manuel*, it makes up less than a quarter of the volume. The rest of the work gives an insight into the more general view that Pareto had of economics.

The first two chapters give his views on the scientific status of the social sciences. He argued strongly that in economics and in the social sciences in general, there were underlying laws and structures which had to be determined, specified and tested by scientific methods. He was contemptuous of the work of his more literary contemporaries and felt that economics should be committed to a positive and minimalist approach. The third chapter is devoted to an introduction to the idea of equilibrium. In chapters 4 and 5, he then separates consumption and production and reduces the individual's problem to one of constrained optimization. In the sixth chapter, he then brings the markets together to talk of the general equilibrium of the system and to discuss its efficiency properties. Two things are worth noting in passing. Pareto was well aware that the important thing for the individual was to maximize subject to the constraints that he perceives, which might or might not be the ones he faces in reality. Secondly, he systematically considered the possibility of monopolistic competition in parallel with that of perfect competition although it is his treatment of the latter that is remembered today.

The chapters following those on economic equilibrium deal with demographic problems and their consequences for the labour force, with many factual illustrations, with natural resources, in particular land, with capital and savings and with monetary problems. All the latter problems are dealt with summarily and consist often of observations based on specific facts. The last chapter is devoted to 'concrete economic phenomena' although these already figure largely in the two preceding chapters.

Ordinal Utility

One of Pareto's major contributions has long been considered as that of establishing that, even though utility may not be measurable, an ordinal notion is sufficient for the construction of equilibrium theory. This important aspect of modern theory was not immediately recognized by Pareto. It was only with his article (1900) that he started to develop a fully ordinal theory. Whether Pareto actually rejected the idea of 'measurable utility' or whether he thought that it was simply impossible to identify the appropriate function is an interesting question. Although he was very clear that any one of the 'indices' of utility would suffice for his analysis, several remarks and a comment of Volterra (1906) lead one to suppose that Pareto had not completely broken with the earlier tradition. Indeed, his discussion of cycles of choice leads him to state that 'In certain cases they would permit knowledge of the value of ophelimity' (*Manuel*, mathematical appendix). Pareto develops in the *Manuel*, his analysis of 'indifference curves' contrasting his approach to that of Edgeworth who started with 'ophelimity' or 'utility' and obtained expressions for the indifference curves' (*Manuel*, p. 540, footnote 1).

In the case of two goods, consider x and y the quantities consumed of those goods and consider the quantities dx and dy which, when added to x and y, leave the consumers' satisfaction unchanged. This gives an equation:

$$f_1[x,y]\mathrm{d}x + f_2[x,y]\mathrm{d}y = 0$$

This equation for the 'indifference line' gives us the expression: d$F[x,y] = 0$ which is satisfied by an infinite number of F, and Volterra (1906) now remarks that 'Ophelimity is one of these functions F'.

Whether or not the notion of the 'true utility function' lingered, as Volterra pointed out, Pareto's treatment required further work if it was to be extended to the case of three or more goods. Pareto's reply to this was his well known 1906 paper on 'Ophelimity in non-closed cycles'. Although this paper has been regarded as an attempt to solve this 'integrability problem', as Chipman (1971) points out, it is really devoted to a treatment of the measurable utility problem.

There is, first of all, as in several of Pareto's works, a preoccupation with the order in which goods are consumed which, to modern eyes, confuses the discussion. Strictly speaking, all goods are dated and hence changing the order of consumption changes the specifications of the bundle of goods in question. The dating convention was not established in Pareto's time so he devotes considerable time to discussing the 'paths of consumption'. Early on (1900) Pareto simply assumed that the order of consumption was the optimal one. However, in his 1906 paper, he argues as follows: if utility depends on the path leading to the final consumption whereas, at the final point, changes in utility resulting from perturbations of that point depend only on the total quantities consumed then, utility is measurable. In fact, as Chipman (1971) points out, in the case discussed by Pareto, utility is independent of the path of consumption. The type of result he was aiming at is closely related to another case already treated by Fisher (1896). In this case, consider the marginal utility ('elementary ophelimity') of a good as dependent only on the quantity consumed of that good. The utility function obtained in this case is measurable, that is, it is invariant up to a linear transformation. Pareto's work here may thus best be regarded as an early attack on the problem of separable utility functions. What is clear is that while Pareto recognized that equilibrium analysis could be carried out using only ordinal utility, he still had not reached the point of abandoning measurable utility as a concept.

The integrability problem mentioned above, that of recovering an underlying utility function from demand behaviour, was already solved in large part by Antonelli (1886) in a paper which Pareto had in his possession, albeit briefly, since he commented on it in a letter to Pantaleoni in 1891. However, Pareto does not seem to have profited from Antonelli's work although Walras had already praised it. Thus for some reason, Pareto did not appreciate what had already been done and did not make any significant contribution in this direction despite assertions in the literature to the contrary.

Returning to the original point, it must be emphasized that Pareto arrived at conditions for economic equilibrium using ordinal utility alone and thus clearly marked out the trail for modern economic theory.

General Equilibrium

Pareto wrote down what are now considered to be the standard equilibrium conditions for the consumer side of economy, showing the equality of the marginal rate of substitution to the price ratio, normalizing the price of money, which he assumed to give direct utility, to one (Manuel, mathematical appendix). Taking these equations:

$$U_x = \frac{U_y}{P_y} = \frac{U_z}{P_z} \cdots$$

together with Walras's Law:

$$(x - x_0) + P_y(y - y_0) + P(z - z_0) \ldots = 0$$

and differentiating he found expressions for

$$\frac{\partial y}{\partial P_y} \cdot \frac{\partial z}{\partial P_z} \text{etc.}$$

These he had already set out in his 1892 article in the *Giornale degli Economisti*. He then shows that if goods are independent i.e.

$$U_{xy} = 0 \quad x \neq y$$

then the conditions $U_{xx} < 0$ $U_{yy} < 0$ etc. imply that demand for each good is a decreasing function of its own price (*Manuel*, Mathematical Appendix, section 53). This is a forerunner of more general but very recent results.

It is worth noting in passing some of Pareto's clarifications in the general context of Marshall's work. Firstly, he showed in the *Manuel* (Mathematical Appendix, section 56) that, in general, the marginal utility of money changes with prices. Thus it cannot arbitrarily be assumed to be constant.

He showed in his Encyclopaedia article (1911, section 23) that if the elasticity of demand for all goods is constant, then it is unity, and remarked that since Marshall had not imposed this restrictive condition his analysis was defective.

Lastly, he showed that the idea of estimating consumer surplus as the area under the consumer's demand curve above the exchange price was wrong unless the marginal utility of money happened to be constant (*Cours*, section 83).

Pareto then gives a numerical example of equilibrium with very special utility functions and a linear production function. The conditions for equilibria with production are dealt with in a much less satisfactory way. Indeed, Pareto's treatment of production has been the subject of much discussion and he has been criticized for having abandoned too easily the 'marginal productivity' approach. As Schumpeter (1954) points out, the case often considered by Pareto is only a limiting case of the standard one and the fact that one has to consider boundary cases does not exclude the appellation 'marginal productivity'.

In fact, much of the difficulty in appreciating Pareto's contribution to equilibrium theory stems from his analysis of production. He dealt with both variable and fixed coefficients in production without separating them clearly, as he might have done, into considerations of the long and short term. He repeated explicitly in the *Cours* (section 714) the hypothesis of constant returns to scale. Furthermore, he abandoned, by his introduction of fixed costs, the convexity of the production set which plays such an important role in later analysis (*Manuel*, ch. 5). Finally, what seems to have escaped attention is the effort that Pareto made to discuss what has come to be called 'monopolistic competition' and its introduction into equilibrium analysis (*Manuel*, chs. 3, 5 and 6). Recognizing that individuals can influence prices and that this should be taken into account led him to try to take explicit account of the demand with which they are faced. This merits two comments. Only recently has the introduction of 'monopolistic competition' into general equilibrium models resurfaced and the article by Negishi (1961), which uses rather arbitrary assumptions, is generally cited as the first example. Thus Pareto was already dealing with a problem which has still not been really satisfactorily treated.

Secondly, the confusion as to the nature of time in all of Pareto's analysis is clear. Indeed, he recognizes this himself, when talking of the passage from an initial position to an equilibrium, when he says (*Manuel*, ch. 3, section 171), that the issues he discusses fall into the domain of dynamics rather than statics. The modern convention that the adjustment process to equilibrium is instantaneous and that long-term dynamics consist of the passage from one equilibrium to another is far from that adopted by Pareto.

Pareto did not make a clear distinction between the question of existence and the question of stability. He regarded equilibrium as the terminating point of a process and this is brought out in the *Cours* and particularly in the *Manuel* (ch. 3, sections 110–15 for example). The time taken for this process was not specified but is certainly not regarded, even conventionally, as negligible. He described the passage or path from the initial position to the final position under assumptions of perfect competition with tâtonnement and non-tâtonnement processes and considered the monopolistic competition case, although rejecting it as too difficult to handle. He did, however, enter into a discussion as to how individuals could push the economy towards a preferred equilibrium in the case of multiple equilibria and showed how they would try to manipulate the terms of exchange along this path (*Manuel*, p. 197) and discussed how individuals would benefit from doing this. Furthermore, in the light of this manipulation he suggested that certain equilibria would be stable. Thus Pareto recognized explicitly that stability is a property of a particular process.

Given this vision of equilibrium, Pareto did not try to show the existence of equilibrium as such except by counting equations and unknowns (*Manuel*, Appendix). In effect, he said that since one could find the conditions for equilibrium, an interesting possibility would be to solve explicitly all the equations necessary to determine the equilibrium. However, as he pointed out (*Manuel*, ch. 3, section 217) 'If we take into account the fabulous number of equations that a population of forty million individuals and several thousand goods would give, it would not be mathematics that would come to the aid of economics, but economics that would come to the aid of mathematics' since such a system would be beyond human capacity to solve. Thus Pareto assumes from a simple argument the existence of a solution and simply dismisses the practicality of finding it for a large economy even if all the relevant equations were known.

Thus the preoccupation with the formal establishment of equilibrium which was later to dominate mathematical economics was not shared by Pareto. Although relatively rigorous, he failed to specify various assumptions such as differentiability and only considered interior solutions to the maximum problem.

Before leaving Pareto's treatment of equilibrium, it is interesting to note that he was clearly aware of the possibility of multiple equilibria and in his diagrams in the *Manuel* (p. 192), he seems to have realized that 'in general, the number of equilibria would be odd', a result proved only very recently.

Pareto also seems to suggest that a socialist state would be better able to lead its economy to an equilibrium than an economy based on private property. The reasoning given for this is based on Pareto's particular view of production and this assertion is also heavily qualified (*Manuel*, ch. 6, sections 58–61). Nevertheless, it is interesting to note this contrast with the view of Pareto as an unqualified liberal.

Efficiency or Pareto Optimality

Of all Pareto's contributions to economics, it is this notion of 'optimality' or efficiency that has made the greatest impact.

Yet it was not he who first gave a definition of a situation corresponding to the modern definition. Edgeworth (1881) clearly defined a situation in which the utility of each individual is maximized given the utilities of the others. Although this definition is given in the context of an exchange economy, its extension to more general cases was not difficult.

It was not so much the introduction of the idea but the use that Pareto made of it which makes his contribution important.

Thus, although he had read Edgeworth, his definition which also includes production is an integral part of his own work.

Pareto defined a notion of surplus or gain which was maximized at an optimum. The real insight that Pareto had was that his notion of efficiency or optimality was independent of all institutional arrangements and of all distributional considerations (*Cours*, vol. 2). Pareto then went on in the *Manuel* (ch. 6 and Mathematical Appendix, sections 145–52), to establish the 'first theorem of welfare economics' that a competitive equilibrium is a Pareto optimum and a tentative version of the 'second theorem' that any Pareto optimum can be obtained as a competitive equilibrium from an appropriate distribution of initial resources. The latter result is only suggested and is never clearly stated. Furthermore, both results are incomplete and even incorrect as a result of the confusion in the treatment of production.

Pareto's ideas on the nature of efficiency evolved over time and in the *Trattato* (sections 2128–39), he showed that the maximization of any social welfare function W which was an increasing function of individual utility functions U_i.

$$W = F(U_1, U_2, \ldots)$$

whether the U_i were defined over the consumption of all individuals or, just restricted to individual consumption gave an optimum. Now as Pareto states (*Trattato*, pp. 1342–3), it is clear that in defining W a government would have to give weights to the different individuals. The idea of including the consumption of other individuals in the utility functions extends the scope of normal economic analysis to what were considered at the time and are still often thought of as 'sociological' considerations.

Pareto did not observe that by appropriately modifying F, all optima could be generated. As Allais (1968) suggests, it is not clear that Pareto was fully aware of the impact of this contribution.

Income Distribution, Pareto's Law

One of Pareto's most remarkable contributions was his development of a 'law' governing the distribution of income. It is significant both as a pioneering piece of applied econometrics and for the controversy that its social implications have aroused. His initial work was published in an article in the Giornale degli Economisti in 1895 then in a memoire (1896) on the 'income distribution curve'. Detailed discussion is given in the Cours (sections 957–65) and in the Manuel (ch. 7, sections 2–31).

Three formulae were proposed by Pareto and the first and most widely cited of these is given by:

$$N(x) = \frac{A}{x^a}$$

where $N(x)$ is the number of people having an income greater than or equal to x. As has been frequently pointed out it has obvious problems where either x tends to zero, or one increases x so that $N(x)$ goes to zero. Nevertheless, Pareto obtained values for α in particular for data for the UK collected by Giffen and obtained for 1843: $\alpha = 1.5$ and for 1879/80: $\alpha = 1.35$. Further computations for Prussia, Saxony, Paris and several Italian cities gave values around 1.5 with a maximum of 1.73. Pareto denied that his 'law' had the status of a physical law, and stated in an article in the *Journal of Political Economy* (1897b) that 'I should not be greatly surprised if some day, a well authenticated exception were discovered.' Nevertheless, he believed that the values of α that he found, α itself being a statistic, were sufficiently close for his law to be 'provisionally accepted as universal'.

Pareto also estimated a distribution of the form:

$$N(x) = \frac{A}{(x+a)^{\alpha}} 10^{-\beta x}$$

where a and β like α are constants. He found a value of β so low that he concluded that a distribution of the form

$$N(x) = \frac{A}{(x+a)^{\alpha}}$$

would suffice.

It is not the universality or otherwise of 'Pareto's law' that has provoked so much controversy, and it has been widely recognized since that other distributions provide more satisfactory fits for particular income data.

It is the relation between 'Pareto's law' and the problem of income inequality that has been the subject of dispute. Pareto says that if the number of individuals with an income over a certain level x in relation to the number of those below that level increases, then inequality diminishes (*Manuel*, ch. 7, section 24). Unfortunately, there was a printer's error in the *Cours* and there, the opposite is stated, although from the footnote (*Cours* Livre III, section 965), it is clear what Pareto intended. There has since been considerable confusion about what Pareto actually said.

Let $N(h)$ be the number of individuals with income above h (the 'minimum income') and $N(x)$ the number above x with $x > h$. Then, as Pareto says, if we define

$$U_x = \frac{N(x)}{N(h)}$$

then, 'income inequality will decrease as U_x increases' (*Cours* Livre III, section 965) (*Manuel*, p. 390, footnote 2).

Allais interprets Pareto as saying the opposite, perhaps following the error in the *Cours*. Yet if we now proceed and assume that 'Pareto's Law' holds, then we have:

$$U_x = \left(\frac{h}{x}\right)^{\alpha}.$$

Since $x > h$ by hypothesis, U_x decreases when α increases and income inequality increases. Allais makes an error in his argument and states that:

$$\frac{N(h)}{N(x)} = \left(\frac{h}{x}\right)^{\alpha}$$

an error identical to that made by Roy (1966). Since both Roy and Allais had started from the original mistake in the *Cours*, this further error should have led them to the same final conclusion as Pareto that income inequality varies in the same direction as α. Roy indeed arrives at this conclusion and contrasts it with the work of Gini and others. Allais made a further error and stated that Pareto believed that income inequality varied inversely with α. All this gives some indication of the sort of confusion that has surrounded Pareto's contribution.

Nevertheless, several remarks are in order. 'Pareto's Law' gives empirically a satisfactory fit for the upper tail of the income distribution (the top 20 per cent according to Lydall 1968) but is clearly inconsistent with the lower end. If α were a constant, then there would be little hope for policies aimed at reducing income inequality, as Pareto pointed out to those in favour of the socialist position. Lastly, Pareto's Law has the peculiar feature that the ratio of the average income above x say $m(x)$ to x itself is a constant given by:

$$\frac{m(x)}{x} = \frac{\alpha}{\alpha - 1}.$$

Allais suggested that this might be taken as Pareto's index of inequality. If this were so, then it would decrease with α the opposite of what Pareto intended.

Economics and the Social Sciences

Pareto's vision of the nature of the social sciences is reflected in his works on sociology (in particular the *Trattato*) and a certain number of his positions mark him out from his contemporaries and his successors. He developed and reinforced his idea that such sciences should be positive and went as far as criticizing his earlier work, taking the 'author' of the *Cours* to task for mixing ethical and positive considerations (*Manuel*, Preface). His defence of positivism was clearly associated with Comte's position (1830) and he was interested in developing a 'positive theory of economic policy'. He argued that 'laws' or relations deduced from specific assumptions should be tested empirically against 'observed statistical laws'. He went further, however, and unlike J.S. Mill (1844) who asserted that to verify hypotheses was not part of the business of science, a position supported by Friedman (1953) and Machlup and others, later, argued that assumptions should be examined to see how reasonable they are (*Trattato*, section 59). The importance of Pareto's statistical work which reflected his standpoint has tended to be overlooked and has been dominated by analysis of his purely theoretical contributions.

His approach to economics reflected a double position. Firstly, he shared Marshall's opinion that economic theory should be aimed at examining 'man as he is' and should not become an abstract intellectual exercise. Secondly, however, while he wished economics to be a relevant science, he condemned attempts to apply too readily economic theory to real problems. He believed that much harm had been done to the cause of 'scientific economics' by such hasty applications.

Finally, it should be remembered that while Pareto was with Weber among the first to expound the principles of 'positive social science', his view of the status of economics was ambiguous. He believed fundamentally, and in this he shared Comte's view, that there should be a universal scientific approach to social science. Yet he recognized the need for, and desirability of specialized disciplines and he was persuaded that economics had a special advantage in that it is by nature a more quantifiable science than the other social sciences. His contributions make it clear, however, that he was not prepared to isolate man's economic activity from his other functions.

Conclusion

Pareto's economic contribution has acquired an increasing reputation over time, unlike his sociological work. Yet as was suggested at the outset, it is disappointing that this reputation should be constructed on the basis of such a small part of his work. His strictly theoretical contributions are an essential part of modern general equilibrium theory. Yet here his education and training pushed him towards an equilibrium notion close to those of classical mechanics and in a certain sense, he helped to lock economics into an unhappily rigid framework.

His more imaginative contributions, his analysis of conflicting interests, his concern with statistical verification and his preoccupation with the place of economics as just one part of a larger structure, have all been left aside.

Thus, paradoxically, Pareto's stature as one of the major economic figures of his time with Walras and Fisher, has been diminished by recent over-emphasis of his most formal contributions.

The Greeks lamented the passing of the universal athlete who was replaced by the almost deformed specialist in each individual sport. Pareto's work should place him clearly as one of the last in the former category but the importance that has been attached to part of his mathematical economics is likely to condemn him unjustly to being described as one of the first of the 'specialists'.

See Also

- ► Edgeworth, Francis Ysidro (1845–1926)
- ► General Equilibrium
- ► Volterra, Vito (1860–1940)
- ► Walras, Léon (1834–1910)

Selected Works

1892–83. Considerazioni sui principii fondamentali dell'economia politica pura. *Giornale degli economisti* 4: 389–420, 485–512; 5: 119–157; 6: 1–37; 7: 279–321.

1893. La mortalità infantile e il costo dell'uomo adulto. *Giornale degli economisti* 7: 451–456.

1894. Teoria matematica dei cambi forestieri. *Giornale degli economisti* 8: 142–173.

1896a. Il modo di figurare i fenomeno economici (A proposito di un libro del dottore Fornasari). *Giornale degli economisti* 12: 75–87.

1896b. La curve della entrate e le osservazioni del professor Edgeworth. *Giornale degli economisti* 13: 439–448.

1896–7. *Cours d'économie politique*, 2 vols. Lausanne: Librairie de l'Université. (Referred to as *Cours* in the text.)

1897a. Aggiunta allo studio curva della entrate. *Giornale degli economisti* 14: 15–26.

1897b. The new theories of economics. *Journal of Political Economy* 5: 485–502.

1899a. Quelques exemples d'application de la méthode de moindres carrés. *Journal de statistique suisse* 121–150.

1899b. Tables pour faciliter l'application de la méthode de moindres carrés. *Journal de statistique suisse* 121–150.

1901a. Sul fenomeno economico. Lettera a Benedetto Croce. *Giornale degli economisti* 22: 131–138.

1901b. Le nuovo teorie economiche. Appunti. *Giornale degli economisti* 23: 235–252.

1902. Di un nuovo errore nello interpretare le teorie dell'economia matematica. *Giornale degli economisti* 25: 401–433.

1906. *Manuale d'economie politica*. Milan: Societa Editrice Libraria. (Referred to as Manuel in the text.) Revised and translated as

Manuel d'économie politique. Paris: Giard et Brière, 1909.

1907–1908. L'interpolazione per la ricerca delle leggi economiche. *Giornale degli economisti* 34: 266–285; 36: 423–453.

1910. Walras. *Economic Journal* 20: 138–139.

1911. Economie mathématique. In *Encyclopédie des sciences mathématiques*, I (iv, 4). Paris: Teubner, Gauthier, Villars. Translated into English as: Mathematical economics. *International Economic Papers* 5: 58–102 (1955). (Referred to as 'Economie mathématique' in the text.)

1913. Il massimo di utilità per una collettività in sociologia. *Giornale degli economisti* 46: 337–338.

1916. *Trattato di Sociologia Generale*, 4 vols. Florence: Barbera. (Referred to as Trattato in the text.) Translated into English and edited by Arthur Livingston as *The mind and society.* New York: Harcourt Brace, 1935.

1918. Economia sperimentale. *Giornale degli economisti* 52: 1–18.

1962. *Lettere a Maffeo Pantaleoni 1890–1923*. Ed. Gabriele De Rosa, 3 vols. Rome: Edizioni di Storia e Letteratura.

All the above works together with a number of other articles, reviews and comments, have been gathered together and published from 1964 on as Pareto, Vilfredo (1964–84) *Oeuvres complets*, sous la direction de G. Busoni, 28 vols. Geneva: Librairie Droz.

Bibliography

Allais, M. 1968. Pareto, Vilfredo: Contributions to economics. In *Encyclopaedia of the social sciences*. New York: Macmillan.

Amoroso, L. 1938. Vilfredo Pareto. *Econometrica* 6: 1–21.

Antonelli, G.B. 1886. *Sulla Teoria Matematica della Economia Politica*. Pisa: Tipografia del Folchetto. Translated by J.S. Chipman and A.P. Kirman as 'On the mathematical theory of political economy' in *Preferences utility and demand*, ed. J.S. Chipman, L. Hurwicz, M.K. Richter, and H.F. Sonnenschein. New York: Harcourt Brace, Jovanovich, 1971.

Atkinson, A. 1975. *The economics of inequality.* Oxford: Oxford University Press.

Barone, E. 1908. Il Ministro della produzione nello stato collettivista. *Giornale degli economisti* 37: 267–293, 391–414.

Borgatta, G. 1924. I rapporti fra la scienza economica e la sociologia nell'opera Paretiana. *Giornale degli economisti* 64.

Borkenau, F. 1936. *Pareto*. London: Chapman & Hall.

Bousquet, G.H. 1927. *Introduction aux systèmes socialistes de Pareto*. Paris: Giard.

Bousquet, G.H. 1928. *The work of Vilfredo Pareto*. Minneapolis: Sociological Press.

Busino, G. 1963. Pareto e le autorità di Losanna. *Giornale degli economisti* NS 22: 260–303.

Cappa, A. 1924. *Vilfredo Pareto*. Turin: Gobetti.

Chipman, J.S. 1971. *Introduction to part II of preferences, utility and demand*, ed. J.S. Chipman, L. Hurwicz, M.K. Richter, and H.F. Sonnenschein. New York: Harcourt Brace, Jovanovich.

Comte, A. 1830. *Cours de philosophie positive*, 6 vols. Paris: Schleicher.

Davis, H.T. 1941. *The analysis of economic time series*. Bloomington: Principia Press.

Demaria, G. 1949. L'oeuvre économique de Vilfredo Pareto. *Revue d'Economie Politique* 59: 517–544.

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

Edgeworth, F.Y. 1896. Supplementary notes on statistics. *Journal of the Royal Statistical Society* 2: 533–534.

Eisermann, G. 1961. *Vilfredo Pareto als Nationalökonom und Soziologe*. Tübingen: Mohr.

Fisher, I. 1896. (A review of) 'La courbe de la répartition de la richesse', by Vilfredo Pareto. *Yale Review* 5: 325–328.

Fossati, E. 1949. Pareto dans son et notre temps. *Revue d'Economie Politique* 59: 585–599.

Frechet, M. 1939. Sur les formules de répartition des revenus. *Revue de l'Institut International de Statistique* 7: 32–38.

Friedman, M. 1953. The methodology of positive economics. In *Essays in positive economics*, ed. M. Freidman, 3–43. Chicago: University of Chicago Press.

Giacolone-Monaco, T. 1957. *Pareto–Walras da un carteggio inedito (1891–1901)*. Padua: Cedam.

Gibrat, R. 1931. *Les inégalités économiques*. Paris: Sirey.

Gide, C. 1917. Le jubilé Vilfredo Pareto. *Revue d'Economie Politique* 31: 426–433.

Johnson, N.O. 1937. The Pareto law. *Review of Economic Statistics* 19: 20–26.

Lydall, H.F. 1968. *The structure of earnings*. Oxford: Oxford University Press.

Machlup, F. 1955. The problem of verification in economics. *Southern Economic Journal* 22: 1–21.

Machlup, F. 1960. Operational concepts and mental constructs in model and theory formation. *Giornale degli economisti e annali di economia* 19: 553–582.

Machlup, F. 1964. Professor Samuelson on theory and realism. *American Economic Review* 54: 733–736.

Mandelbrot, B. 1963. New methods in statistical economics. *Journal of Political Economy* 71: 421–440.

Marshall, A. 1920. *Principles of economics*, 8th ed. London: Macmillan (first published in 1890).

Mill, J.S. 1844. *Essays on some unsettled questions of political economy*. London: J.W. Parker.
Mortara, G. 1924. Pareto statistico. *Giornale degli economisti* 64: 120–125.
Negishi, T. 1961. Monopolistic competition and general equilibrium. *Review of Economic Studies* 28: 196–201.
Pantaleoni, M. 1907–1908. *Pure economics*. New York: Kelly and Macmillan, 1957. Originally published in Italian as 'Elementi di economia pura', and translated into English in 1898. *Giornale degli economisti* 34: 266–285; 36: 423–453.
Pantaleoni, M. 1923. Vilfredo Pareto. *Economic Journal* 33: 582–590.
Pantaleoni, M. 1924. In occasione della morte di Pareto reflessioni. *Giornale degli economisti* 44: 1–19.
Pirou, G. 1938. *Les théories de l'équilibre économique': Walras et Pareto*. Rome: Giornale degli Economisti e Rivista de Statistica.
Roy, R. 1966. Preface to V. Pareto, *Statistique et économie mathématique*. Geneva: Librairie Droz.
Samuelson, P.A. 1963. Problems of methodology – Discussion. *American Economic Review: Papers and Proceedings* 53: 231–236.
Samuelson, P.A. 1964. Theory and realism: A reply. *American Economic Review* 54: 736–739.
Schumpeter, J. 1949. Vilfredo Pareto (1848–1923). *Quarterly Journal of Economics* 63: 147–173.
Schumpeter, J. 1949, 1965. Vilfredo Pareto. In *Ten great economists, from Marx to Keynes*, ed. J.A. Schumpeter, 110–142. New York: Oxford University Press.
Schumpeter, J. 1954. *History of economic analysis*. New York: Oxford University Press.
Spengler, J.J. 1944. Pareto on population 1. *Quarterly Journal of Economics* 58: 593–598.
Stark, W. 1963. In search of the true Pareto. *British Journal of Sociology* 14: 103–112.
Volterra, V. 1906. *L'Economia matematica e il Nuovo manuale* del Prof. Pareto. *Giornale degli economisti* 32: 296–301. Translated by A.P. Kirman, revised and edited by J.S. Chipman as 'Mathematical economics and Professor Pareto's new manual', in *Preferences utility and demand*, ed. J.S. Chipman, L. Hurwicz, M.K. Richter, and H.F. Sonnenschein. New York: Harcourt Brace, Jovanovich, 1971.
Walras, L. 1965. *Correspondence of Léon Walras and related papers*, 3 vols, ed. W. Jaffé. Amsterdam: North-Holland.
Weber, M. 1913. *Wirtschaft und Gesellschaft*. Originally printed as the third part of *Grundriss der Sozialökonomik*, 2 vols. Tübingen: Mohr. Reprinted in *Gesammelte Aufsätze zur Wissenschaftslehre*. Tübingen: Mohr, 1922.
Weber, M. 1949. *The methodology of the social sciences*. Translated and edited by E. Shils and H. Finch. Glencoe: The Free Press.
Wicksell, K. 1897. Vilfredo Pareto, Cours d'économie politique. *Zeitschrift für Volkswirtschaft, Sozialpolitik und Verwaltung* 159–166. Reprinted in English in *Selected papers of Knut Wicksell*, ed. E. Lindahl, 141–158. Cambridge, MA: Harvard University Press, 1958.

Pareto Distribution

Josef Steindl

JEL Classifications
D3

Using certain data on personal income V. Pareto (1897) plotted income on the abscissa and the number of people who received more than that on the ordinate of logarithmic paper and found a roughly linear relation. This Pareto distribution or Pareto law may be written as

$$x = ay^{-\alpha} \text{ or } \log x = a' - \alpha \log y \qquad (1)$$

where α (the negative slope of the straight line) is called the Pareto coefficient. The density of the distribution is

$$dx = a\alpha y^{-\alpha+1} dy$$

The Pareto coefficient is occasionally used as a measure of inequality: The larger α the less unequal is the distribution. According to Champernowne (1952), α is useful as a measure of inequality for the high income range whereas for medium and low incomes other measures are preferable.

α takes only positive values. If $\alpha < 2$ the distribution has no variance; if $\alpha < 1$ it has no mean either. In practice the Pareto law applies only to the tail of the empirical distributions i.e. to incomes above a certain size. Thus the law (1) is valid asymptotically as $y \to \infty$. The range in which the empirical distributions conform to the law is different in different cases. It seems to be larger for wealth than for income (perhaps because we have data only for large wealth) and even larger for towns. In the case of firm sizes only very large firms are covered by the law.

In the case of the distribution of towns by size of population the rank-size relation has been used (Zipf 1949) which is the same as the Pareto distribution except that it uses rank as a measure of the tail (instead of the number of towns above a certain size) so that the higher the rank (beginning with rank one for the largest town) the smaller the size of the town. Zipf believed (incorrectly) that the coefficient α is always about one so that the product of rank and size is constant. But Pareto, of course, was even more 'out' with his belief that the Pareto coefficient for income α always equals unity. In highly industrialized countries today it is above 2 and sometimes above 3.

The main interest of the Pareto distribution lies not in its rather limited use as a measure of inequality but in the explanations it has provoked, naturally so since regular patterns are felt to be a challenge to the mind. There are two types of approach to the problem. That of Champernowne, Yule and Simon explains the characteristic pattern as the steady state of a stochastic process which has been evolving in time, so that the pattern reflects something which has been going on in the past. In contrast, Mandelbrot has been looking for a 'synchronic' explanation which does not depend on a process in time. He is mainly concerned with the reproductive quality of the Pareto distribution: If a large number of independent random variables is identically distributed according to Pareto's law then the sum of these random variables will also be distributed according to this law. Thus it could be expected that the income of the various counties in England would be Pareto distributed because it results in each case from the addition of individual incomes which are Pareto distributed.

Champernowne's pioneering work (1953) in essence goes back to his fellowship dissertation of 1936, published in 1973. He builds on a tradition which explains the normal distribution as the result of the addition of random unit steps (left or right) on the line over a long time (random walk; for the terms and concepts relating to random processes, see Feller, Vol. I). If the random walk takes place on the logarithmic scale the distribution of the sum of steps will tend to log normality. This does not give, however, a stable distribution, because the dispersion will go on increasing all the time. Champernowne chooses the technique of the Markov chain: Each year's income depends only on the previous year's income plus a random increment proportionate to last year's income; the probability of various increments remains constant from one year to the other. This feature is called the law of proportionate effect. Thus the required data will be embodied in a matrix which contains the probabilities of transition from one income in one year to another income in the following year. The number of income receivers remains stable in Champernowne's model because each exit is assumed to be automatically compensated by a new entry. To guarantee that the system reaches a steady state it is assumed that on the average the change of income is downwards; this is necessary to compensate the tendency of the system to diffusion which is characteristic of the unrestrained random walk. The assumption reflects the low income of new entrants.

In fact the role of new entry is crucial not only in this model but in other applications as well (size of firms, towns, wealth).

H. Simon (1955) studied the number of times a particular word (vocable) occurs in a text. The number of vocables which occur with a given frequency decreases with that frequency in a Pareto-like fashion. Simon's treatment is based on the work of Yule (1924), who dealt with a biological problem: the frequency of genera with different number of species which is distributed according to Pareto. He explained this pattern by means of a pure birth process deriving from this the Yule distribution with density

$$f(n) = \alpha\Gamma(1+\alpha)n^{-1-\alpha} \text{ as } n \to \infty.$$

The model of evolution assumes that mutations occur randomly with a frequency g per time unit, creating new genera, and with a frequency s per time unit creating new species, where $g < s$. Since each species has the same chance of creating a new species we have here a proportionate growth, in analogy to the law of proportionate effect. The steady state is produced by the emergence of new genera. The Pareto coefficient equals the ratio of the frequencies with which the two kinds of mutations appear, that is g/s. Simon, whose merit it is to have drawn attention to this brilliant work, has suggested application to incomes (not very convincingly) and has himself applied it to firm sizes (1967). A very direct application relates to the size of towns (Steindl 1965). If the number of towns grows at the rate of μ and the number of inhabitants of the town grows at the rate of ρ then after a sufficiently long time there will be a steady state distribution with Pareto coefficient μ/ρ.

Mandelbrot (1960, 1961) deals with the problem from the point of view of a mathematician and therefore on a very general level. He starts from the concept of stable laws (compare Feller, Vol. II, ch. VI). If a sum of independent identically distributed random variables is distributed in the same way as its components, except for a scale factor and possibly of a location factor, then this distribution is stable. The best-known example is the normal distribution. It has been shown by *P.* Lévy that there is a class of distributions with infinite variance which are stable and which converge to the law of Pareto when the variable in question (say, income) tends to infinity. The Pareto law in this context is confined to the range $1 < \alpha < 2$. Mandelbrot surmises, owing to the reproductive quality, in the above sense, of the Pareto law, that its importance empirically must be very great. He also considers that this must have implications for some statistical methods which depend on the assumption of normalcy.

As to income, Mandelbrot suggests that it can be regarded as composed of a number of independent elements which are identically distributed. We can easily imagine a decomposition into a few parts such as earned income, property income and transfer income. Mandelbrot requires, however, in order to assure convergence, a large number of components, and these, as he admits, have hardly any counterparts in reality (1961, p. 525). The explanation is analogous to the well known explanation of the stature of adult men as a random variable composed of a great number of independent small random variables; this explains the normal distribution of height. The precise identity of these small random variables is, here again, not specified and rather speculative. This may perhaps explain why this ‘synchronous’ approach has not, so far, found much resonance among economists.

The interest of the alternative approach (Champernowne or Yule) of explaining the law as a steady state of a stochastic process is that it establishes a relation between the stratification found in a cross section and the past history which has produced it, and which is mapped in the cross section. This is analogous to the stratifications in geology or the rings in the trunk of a tree. Irregularities or shifts in the empirical distributions can according to this view be explained by major disturbances of the process in certain points of time in the past.

Concretely, the Pareto distribution has been shown, in the case of a birth and death process model, to depend on growth; in an economy which has always been stationary it would not exist (Steindl 1965). The Pareto coefficient in such models is usually a ratio of growth rates; thus in the case of firm size it is a ratio of the growth rate of the number of firms to the growth rate of the firms themselves (Steindl 1965). The importance of new entry as a factor making for less inequality has also been shown, *inter alia* in the case of wealth (Steindl 1972).

The stochastic models have often been criticized for their lack of economic content. Perhaps it has been overlooked that they only represent the first steps in a new and exceedingly difficult terrain. It may be thought that the work of Champernowne, Yule, Simon and Wold and Whittle contains the seed of future studies which will reveal their full potentiality only when they are extended to distributions in several dimensions.

See Also

- ▶ Gini Ratio
- ▶ Lognormal Distribution
- ▶ Lorenz Curve

Bibliography

Champernowne, D.G. 1952. The graduation of income distributions. *Econometrica* 20: 591–615.

Champernowne, D.G. 1953. A model of income distribution. *Economic Journal* 63: 318–351. Reprinted in D.-G. Champernowne. 1973. *The distribution of income between persons.* Cambridge: Cambridge University Press.

Feller, W. 1950, 1966. *An introduction to probability theory and its applications.* 2 vols. Reprinted, New York: John Wiley & Sons, 1968, 1971.

Ijiri, Y., and H.A. Simon. 1964. Business firm growth and size. *American Economic Review* 54: 77–89.

Mandelbrot, B. 1960. The Pareto–Lévy law and the distribution of income. *International Economic Review* 1 (2): 79–106.

Mandelbrot, B. 1961. Stable Paretian random functions and the multiplicative variation of income. *Econometrica* 29 (4): 517–543.

Pareto, V. 1897. *Cours d'économie politique*. Lausanne: Rouge.

Simon, H.A. 1955. On a class of skew distribution functions. *Biometrika* 42: 425–440. Reprinted in H.A. Simon. 1957. *Models of man: Social and rational.* New York: John Wiley.

Steindl, J. 1965. *Random processes and the growth of firms. A study of the Pareto law.* London: Griffin.

Steindl, J. 1972. The distribution of wealth after a model of Wold and Whittle. *Review of Economic Studies* 39 (3): 263–279.

Wold, H.O.A., and P. Whittle. 1957. A model explaining the Pareto distribution of wealth. *Econometrica* 25: 591–595.

Yule, G.U. 1924. A mathematical theory of evolution based on the conclusions of Dr. J.C. Willis. *Philosophical Transactions of the Royal Society of London Series B* 213: 21–87.

Zipf, G.K. 1949. *Human behavior and the principle of least effort*. Reading: Addison-Wesley.

Pareto Efficiency

B. Lockwood

Keywords

Bounded rationality; Constrained efficiency; Cost–benefit analysis; Efficiency; Externalities; Incomplete contracts; Incomplete markets; Lump sum taxes; Market failure; Market socialism; Multiple equilibria; Neoclassical economics; Optimal taxation; Pareto efficiency; Pareto optimality; Pareto, V.; Planning; Signalling; Socialist calculation debate; Welfare economics; X-efficiency

JEL Classifications

D6

Few concepts are more widely used within economics than that of 'efficiency'. It usually means not wasteful, or doing the 'best' one can with available resources. However, there are specialized usages; for example, the concept of efficient markets in the finance literature, or Leibenstein's concept of X-inefficiency. Not all these meanings, even in academic work, have a common provenance. However, the concept as used in neoclassical economics has a precise but rather narrower meaning, given to it by Pareto, the Italian economist and sociologist, in his works *Course in Political Economy* and *Manual of Political Economy* around the turn of the twentieth century. He suggested the following definition: an allocation of resources in the economy was optimal if there existed no other productively feasible allocation which made all individuals in the economy at least as well-off, and at least one strictly better off, than they were initially. Although Pareto actually used the word 'optimal', this is really a definition of efficiency, as a Pareto-'optimal' allocation of resources is 'good' only in the limited sense that not everybody can be made better off. It may in fact be very undesirable in some other way, for example, very unequal. It is not surprising, therefore, that the word 'Pareto-optimal' has gradually been replaced by 'Pareto-efficient'.

There are several points to note about this definition. First, it is only well defined within a neoclassical framework, that is, where the preferences of individuals and the technical possibilities of production are taken as the ultimate data of economic analysis. Secondly, even within this framework, it is an ordinal concept of efficiency, as it does not rely on any intensity of preference, interpersonal comparability of utilities, or commensurability of different inputs or outputs for its definition. This is no accident; Pareto was a convinced ordinalist, who believed that the utilitarian concept of introspective utility was unscientific (see for example Pareto 1927, p. 113). Thirdly, while it provides a ranking of allocations of economic goods between individuals, it does not permit a ranking of all such allocations, that is, there are many different allocations that are Pareto-optimal and which differ with respect to the distribution of real income (that is, utility) among the individuals in society.

Simple and limited idea though this is, it has had an enormous influence on the development of neoclassical economics. First and foremost this is because Pareto did not simply present this notion of optimality as an abstract criterion, but showed that competitive equilibrium would yield an optimal allocation of resources in this sense, thus making precise the notion of the 'invisible hand'. It is no exaggeration to say that the entire modern microeconomic theory of government policy intervention in the economy (including cost–benefit analysis) is predicated on this idea. It also stimulated other debates, such as the one over 'market socialism' in the 1930s, which led to modern theories of economic planning. However, it has by its very success inhibited investigation of other criteria for the performance of economic systems. More radical commentators argue that Pareto, and what followed after, also serves an ideological purpose, namely, to show that capitalism is inherently self-regulating, with phenomena such as unemployment being explained as deviations from an 'ideal' equilibrium rather than inherent structural problems. In this article, we briefly review the historical evolution of the idea, and then attempt a critical assessment.

As already remarked above, the context in which Pareto first presented his concept of optimality was in demonstrating that competitive equilibrium was optimal, of efficient. This crystallized the notion, present at least since Adam Smith, that free trade has (possibly unintended) beneficial consequences. His arguments were much refined and extended over the years by figures such as Barone, Lerner, Hicks and Samuelson, although it took some 20 or 30 years after the *Manual of Political Economy* was published for the 'new' welfare economics to become common currency. The current version of the proposition is essentially based on the work of Arrow and Debreu (for example, Debreu 1959) who generalized and clarified the mathematics of the result. They showed that it is in fact a twofold proposition; under certain conditions, competitive equilibrium is Pareto-efficient, and second, the additional assumption of non-increasing returns to scale, any Pareto efficient allocation of resources may be decentralized as a competitive equilibrium. These statements are known collectively as the two theorems of welfare economics.

Before going on to discuss them, one should note that there is, to begin with, a problem with the notion of 'competitive'. By this, we simply mean here that all firms (or more generally, all agents) take prices as given, not necessarily that they are 'small' relative to the economy. The problem is that the former is not generally plausible behaviour unless the latter is true. Therefore, the result should really be thought of as approximate – that is, with price or quantity-setting firms equilibrium is approximately competitive, and hence approximately optimal when firms are 'small' – although it is not usually presented in this way.

Now, given price-taking behaviour, the sufficient conditions for the first theorem are (i) that there are no externalities and (ii) that there are complete contingent markets for all commodities (apart from externalities), that is, markets at all present and future dates and states in all contingencies. Implicit in (ii) is the assumption that all agents are equally and perfectly informed about all aspects of their environment. The reason why the first condition is sufficient (and, generally, necessary) is simply that externalities such as air pollution, are in this framework goods (or, more properly, 'bads') for which no markets exist, so there is no mechanism for the marginal benefits of the externality-producing activity to be equated to the marginal damages they impose on others.

The role of complete contingent markets, however, is not so immediately apparent. The reason is as follows. Consider, for example, a two-period economy with spot markets, but no means of transferring income from period to period (that is, no securities or money). Then, clearly, individual marginal utilities of income will not be equalized in the two periods. Equalization is, however, a necessary condition for full Pareto efficiency, as otherwise a reallocation of income between periods can make at least one agent better off than in the original position. The same argument applies *a fortiori* if there are no, or limited, means of transferring income between different states of the world. If there are complete contingent markets, however, this problem cannot arise.

Some, however, have suggested that, with incomplete markets, full Pareto efficiency is too demanding a performance criterion, and have suggested that one should reformulate the concept to take into account the inherent restrictions on allocation of resources when markets and information are incomplete. This is sometimes called constrained Pareto efficiency. The problem with such an approach is that the exact definition of constrained efficiency is often arbitrary. To take an example, Hart (1975) proposed the definition that a competitive equilibrium with incomplete markets was constrained Pareto-optimal if there was no other competitive equilibrium relative to the same allocation of endowments which Pareto-dominated it. This seems a weak criterion, (for a start, it only has force when there are multiple equilibria), but nevertheless he showed surprisingly that not all equilibria were Pareto-efficient even in this sense, that is, that multiple equilibria could be Pareto-ranked. On the other hand Gale (1982) has proposed an even weaker notion of Pareto efficiency relative to which the first theorem is true, and there is no way of deciding which 'the' correct measure of performance is. In addition, the issue becomes even more complex in the more interesting case where information is asymmetric, with the concomitant phenomena of signalling, adverse selection, moral hazard, and so on. Here, competitive equilibrium may 'fail' in a number of new ways; for example, resources may be wasted on signalling. In summary, Pareto's argument is not general; the invisible hand becomes very shaky when unrealistic assumptions are dropped.

Therefore, very few people take the theorems of welfare economics seriously as descriptions of the real world. The main significance of the two theorems has been in generating a framework for evaluation of government intervention in the economy: this framework has dominated neoclassical thinking about public policy. One can distinguish two types of policy analysis. The first, which we can call 'market failure' analysis, abstracts from distributional considerations by supposing that the government can lump-sum tax individuals in the economy. The procedure is to compare the 'real' economy to the complete contingent markets economy, which is known to be efficient, and on the basis of this prescribe policies that either mimic or replace markets to some extent, or more generally alleviate the inefficiency. The classic example of this approach is in the externalities literature, which proposes either the creation of artificial markets in the externalities via the assignment of property rights (à la Coase), or the imposition of corrective taxes (à la Pigou). This kind of prescriptive analysis may seem excessively utopian; however, some have used the market failure paradigm descriptively to explain the existence of various institutions as replacements for markets, a classic example being Arrow's (1963) discussion of health care.

The second type of policy analysis is concerned with the 'problem of redistribution' or how to redistribute pre-tax incomes to satisfy distributional objectives without the benefit of lump-sum taxation, that is, using income or commodity taxes, and so on. In this case, the first theorem says that the initial no-tax equilibrium is efficient, but, given that redistribution involves distortionary taxation, with redistributive taxes competitive equilibrium will be Pareto-inefficient. Hence, there is a trade-off between Pareto-efficiency and distributional objectives. The literature has, in the main, been concerned with characterizing those tax structures that are 'second best' Pareto-efficient, that is, such that there is no change in the available taxes such that all agents can be made better off. There is now a large literature on such optimal redistributive taxation (see for example Mirrlees 1986).

There are, of course, major problems with the actual implementation of the policy prescriptions that arise from this analysis. First, usually the policy recommendations depend on the taste/technology specification of the models (for example, optimal taxation formulae depend crucially on the functional forms chosen for labour supply and commodity demands) and the latter are only testable to a very limited extent. Second, as Lipsey and Lancaster (1956) showed, the 'optimal' policies are also generally sensitive to assumptions made about the existing 'distortions' in the economy (for example, taxes, monopolies, and so on) if these are not also controllable by the government or planner. For, as they showed, it may not be desirable to substitute lump-sum taxation for a tax on one commodity if there is a preexisting tax on another. Finally, the characterization of the trade-off between Pareto efficiency and distribution needs to be complemented by a distributional judgement to provide concrete policy recommendations, for example, a rate of income tax.

The alternative approach to policy, of course, is to suppose that the planner is interested in pursuing a number of 'intermediate' objectives such as maximizing the growth of national income or employment, or reduction of inequality of income, or inflation, or some weighted combination of these. However, while these objectives are undoubtedly more operational, and perhaps more philosophically appealing as they do not commit one to a utility-based view of welfare, the above problems will still arise with intermediate objectives.

Therefore, the Paretian approach to policy may have a role to play, especially (i) when Pareto-efficient policies are relatively robust to the structure of tastes/technology and distributional goals as are for example some shadow-price rules for cost–benefit analysis (see for example Drèze and Stern 1987, pp. 49–62) and (ii) to critically analyse the basis for the choice of intermediate objectives; for example, why is inflation 'costly'? When Pareto presented his original proof of the efficiency of competitive equilibrium, he seemed also to assert that these conditions could only be attained in a decentralized economy; 'if one could know all these equations (which describe the optimum) the only means to solve them which is available to human powers is to observe the practical solution given by the market' (Pareto 1927). Nevertheless, Barone pointed out explicitly shortly afterwards that the same efficient allocation of resources could be achieved by an omniscient central planner in a 'socialist' economy, that is, one where the means of production were collectively owned. This, and other subsequent contributions provoked a debate between, among others, von Mises, Lange and Hayek (see for example Lange 1936, or von Hayek 1940) about how – if at all – in practice the Central Planning Board (as Lange called it) could achieve this. Lange, for example, proposed a price-based iterative procedure where the CPB effectively replaced the Walrasian auctioneer. Other solutions were also proposed, and since the 1950s these have been extended and formalized by Arrow, Hurwicz, Malinvaud, Heal and others (see for example Heal 1986). All these schemes, however, essentially use the price system as a means of transmitting information to the central planner. In the end, though, it is questionable whether the market socialism debate has had any real impact on the adoption of market socialism in centrally planned economies.

The concept of Pareto optimality, while simple, almost trivial in itself, has had an enormous impact on economics. However, by providing an apparently precise measure of the efficiency of an economic system, independent of distributional questions, it has, in my view, inhibited discussion of distributional questions and alternative criteria of efficiency. One reason for this, however, may be that within the ordinal neoclassical paradigm Pareto's definition is the only tenable concept of efficiency; that is, any other concept of efficiency, once reformulated in a neoclassical model, will eventually reduce to it.

An example of this is Leibenstein's (1966) notion of X-inefficiency. When Leibenstein introduced the idea, he sharply distinguished it from the notion of Pareto efficiency. The former was inefficiency in the process of production due to the fact that contracts for labour are incompletely specified, the 'production function' is not known, and so on, and so derived from bounded rationality. However, Hart (1983) attempted to capture the notion of X-inefficiency in a fully rational, maximizing model; he identified it with the loss of output due to the fact that managers' efforts cannot be perfectly observed by shareholders. In this framework, X-inefficiency reduces to Pareto inefficiency relative to the full-information equilibrium. A similar fate befalls Schumpeter's definition of efficiency (see Schumpeter 1942, p. 188), which emphasizes the long-run performance of the economy – capital accumulation, technical progress, and so forth. He proposed that perfect competition, which is Pareto-efficient in a static sense, would not be efficient in this long-run sense compared with monopolized industries, as the latter survive better in the 'gale of creative destruction', as he describes capitalism. However, this concept of long-run efficiency reduces to Pareto efficiency if one writes down a dynamic model of competition. This is not to say that Schumpeter's ideas can all be adequately modelled in a neoclassical framework – they cannot – but simply that no other concept of efficiency can sensibly be formulated within this framework.

Therefore, Pareto efficiency and the neoclassical paradigm go hand in hand. If one rejects some aspects of the paradigm, then Pareto efficiency may not have much meaning. For example, some (for example, Galbraith) have argued that, in practice, there is little consumer sovereignty; if desires are manipulated and fears exploited by advertising and so on, the Pareto criterion is of little use in gauging how well real wants are satisfied. Some Marxists (see for example Rowthorn 1980) go further than this, and argue that Pareto's proof of optimality serves an ideological purpose, by presenting a picture of capitalism as a harmonious enterprise and distracting attention from its exploitative nature.

See Also

- ▶ Optimality and Efficiency
- ▶ Rational Behaviour
- ▶ Welfare Economics

Bibliography

Arrow, K.J. 1963. Uncertainty and the welfare economics of medical care. *American Economic Review* 53: 941–973.

Debreu, G. 1959. *The theory of value*. New York: Wiley.

Drèze, J., and N. Stern. 1987. The theory of cost–benefit analysis. In *Handbook of public economics*, ed. A. Auerbach and M. Feldstein, vol. 2. Amsterdam: North-Holland.

Gale, D. 1982. *Money: In equilibrium*. Cambridge: Cambridge University Press.

Hart, O. 1975. On the optimality of equilibrium when the market structure is incomplete. *Journal of Economic Theory* 11: 418–443.

Hart, O. 1983. The market mechanism as an incentive scheme. *Bell Journal of Economics* 14: 366–382.

Heal, G. 1986. Planning. In *Handbook of mathematical economics*, ed. K. Arrow and M. Intriligator, vol. 3. Amsterdam: North Holland.

Lange, O. 1936. On the economic theory of socialism, part 1. *Review of Economic Studies* 4 (1): 53–71.

Leibenstein, H. 1966. Allocative efficiency v. X-efficiency. *American Economic Review* 56: 392–415.

Lipsey, R.G., and K. Lancaster. 1956. The general theory of second best. *Review of Economic Studies* 24: 11–32.

Mirrlees, J. 1986. The theory of optimal taxation. In *Handbook of mathematical economics*, ed. K. Arrow and M. Intriligator, vol. 3. Amsterdam: North-Holland.

Pareto, V. 1927. *Manual of political economy*, 1971. London: Macmillan.

Rowthorn, B. 1980. Neo-classicism, neo-Ricardianism, and Marxism. In *Capitalism, conflict and inflation*, ed. B. Rowthorn. London: Lawrence & Wishart.

Schumpeter, J.A. 1942. *Capitalism, socialism, and democracy*. London: Allen & Unwin.

von Hayek, F.A. 1940. Socialist calculation; The competitive solution. *Economica* 7: 125–149.

Pareto Principle and Competing Principles

Louis Kaplow

Abstract

The Pareto principle, the seemingly incontrovertible dictum that if all individuals prefer some regime to another then so should society, may conflict with competing principles. Arrow's impossibility theorem and Sen's liberal paradox are two notable examples. Subsequent work indicates more broadly that the Pareto principle conflicts with all non-welfarist principles. This essay surveys these results, including various extensions thereof, and offers perspectives on the conflict, drawing on classical and contemporary work in political economy and economic psychology.

Keywords

Arrow, K.; Arrow's impossibility theorem; Behavioural economics; Bentham, J.; Darwin, C.; Harrod, R.; Harsanyi, J.; Hume, D.;

Interpersonal utility comparisons; Liberal paradox; Mill, J. S.; Pareto principle; Philosophy and economics; Psychology and economics; Rawls, J.; Robbins, L.; Second best; Sen, A.; Sidgwick, H.; Smith, A.; Social choice; Social welfare function; Welfare economics

JEL Classifications
D63; H43

The Pareto (1906) principle holds that, if all individuals strictly prefer one state, regime, or policy to another, then that selection is deemed socially preferable as well. Because of the power of unanimous endorsement, the Pareto principle has understandably been important in normative economic analysis. Even though strict Pareto dominance is unlikely to prevail when society is deciding among plausible competing alternatives (for this would require that literally each of millions preferred the same outcome), the Pareto principle nevertheless offers important guidance. In particular, the principle may help in choosing among or ruling out various other evaluative notions; principles that turn out to conflict with the Pareto principle may accordingly be rejected. Alternatively, if some competing principles seem compelling, they may raise doubts about the ostensibly incontrovertible Pareto principle.

The first sections to follow review two well-established conflicts between the Pareto principle and certain competing principles: Arrow's (1951) impossibility theorem and Sen's (1970) liberal paradox. The succeeding section presents more recent work that establishes a general conflict between the Pareto principle and all non-welfarist notions, whether they concern rights, justice, or other conceptions of fairness (apart from those pertaining only to the distribution of welfare itself). A final section examines classically grounded strands of literature on political economy and economic psychology that help reconcile the tension between the seemingly unimpeachable Pareto principle and conflicting non-welfarist principles, many of which have appeal to the public, policymakers, and economists as well. (The Pareto principle is also important in normative economic analysis, notably with regard to the two fundamental theorems of welfare economics, a subject not considered in this article.)

Arrow's Impossibility Theorem

Perhaps the most famous instance of conflict between the Pareto principle and competing principles is Arrow's (1951) impossibility theorem. Arrow considered social choice procedures designed to generate a consistent social ordering (a complete and transitive ranking) from purely ordinal information about individuals' preferences. In one formulation of Arrow's theorem, the assumptions of universal domain (no restriction on individuals' preferences), independence of irrelevant alternatives (the social ordering of any two alternatives depends only on individuals' orderings of those two alternatives), non-dictatorship (no one individual's preferences completely determine social preferences), and the Pareto principle imply that such a social ordering is impossible.

A large subsequent literature explores whether relaxing some of Arrow's assumptions modestly would make possible procedures that yield robust social orderings. Of particular relevance here are attempts to weaken the Pareto principle. As surveyed in Campbell and Kelly (2002), these efforts have been largely unsuccessful: either there are frequent violations of the Pareto principle or a single individual will have substantial, even if not completely dictatorial, influence.

Nevertheless, Arrow's theorem does not rule out the class of standard, individualistic social welfare functions (SWFs), mappings from individuals' utilities to a measure of social welfare, that are fully consistent with the Pareto principle. Consider the discrete case, in which there are n individuals, $U_i(x)$ is the utility of the ith individual, and x is a complete description of the pertinent state. Then we can define $W(\mathrm{U}_1(x),\ldots,U_n(x))$ as an individualistic SWF (so called because it depends only on individuals' utilities). Assuming, as is standard, that W is increasing in each individual's utility, it follows that, for any set of individuals' utility functions $\{U_i(x)\}$, W provides a complete and transitive social ordering of all possible social states that is independent of irrelevant alternatives, non-dictatorial, and satisfies the Pareto principle. The classical utilitarian criterion, $W = \Sigma U_i(x)$, is an example of such an SWF.

The possibility of an SWF is restored by altering Arrow's framework to allow the domain of social choice procedures to consist of individuals' utilities rather than just their orderings. This approach entails interpersonal utility comparisons, which during the mid-20th century (and to an extent thereafter) were eschewed in welfare economics, following the argument of Robbins. As Robbins (1935, vii–x, 1938) himself clarified in the second edition of *An Essay on the Nature and Significance of Economic Science* and a subsequent essay, however, his argument was not that interpersonal comparisons should not be made – indeed, they were inevitable – but rather that they involve value judgements rather than scientifically verifiable statements. Much modern welfare economics has pursued analysis of SWFs that depend on individuals' utilities and not just orderings, presumably because of a belief that preference intensities matter and that interpersonal comparisons are required if distributive judgements are to be made.

Sen's Liberal Paradox

In 'The impossibility of a Paretian liberal', Sen considered whether the Pareto principle conflicts with a specific notion of liberalism, subsequently described by many (including, on occasion, Sen himself) as a species of libertarianism. His condition stipulates that there exist certain choices about which the social ranking should reflect that of a particular individual, regardless of other considerations, including effects on the utility of others. This conception and Sen's analysis thereof is well illustrated by considering his much-discussed example. One individual, whom we shall call Prude, abhors erotic literature, and a second, Lewd, adores it. Both individuals' preferences, moreover, are assumed to be meddlesome in the following manner. Prude would be more upset by Lewd's reading a certain lascivious novel than reading it himself, and Lewd would get more pleasure from Prude's reading the novel than reading it herself. Therefore, as between just Prude reading the novel and just Lewd reading it, both prefer the former. However, Sen's liberal principle insists that the latter be the social choice: Prude's preference against his own reading of the book, ceteris paribus, dictates socially that Prude should not read the book, and likewise Lewd's desire that she read the book, ceteris paribus, dictates socially that Lewd should read it. Hence, the choice that Sen's liberal principle deems socially best is one that would be rejected under the Pareto principle.

Analytically, Sen's result can be understood by reference to the familiar concept of externalities. Lewd's reading the book involves a negative externality on Prude, whereas Prude's reading the book involves a positive externality on Lewd. (Compare the case in which Lewd moderately enjoys loud parties that greatly annoy his neighbour Prude, and Prude would rather not bother to replace his weed-ridden garden with flowers that would greatly delight his neighbour Lewd.) Failing to regulate externalities obviously may violate the Pareto criterion. Furthermore, in Sen's example, the two individuals – if left to themselves – would wish to enter a Coasian bargain under which Prude, rather than Lewd, reads the book (just as, in the variation, Lewd should agree to refrain from loud parties if Prude agrees to replace his weeds with flowers). Sen's principle implicitly prohibits both government regulation and private exchange in which individuals mutually relinquish their posited liberal rights. Preventing mutual waiver both by vote and by contract may hardly seem liberal, as argued by Gibbard (1974) and many others in a highly elaborated literature, surveyed by Suzumura (forthcoming). Indeed, any notion that conflicts with the Pareto principle must embody an underlying opposition to freedom since a violation of the Pareto principle entails contravention of unanimous choice. Some of Sen's subsequent writing (for example, 1992, pp. 144–6) defends his original liberal principle on grounds of practicality and concern for governmental abuse of power. As explored below in the final section, however, such Millian (Mill 1859) justifications for rights may be powerful but are not, at root, inconsistent with the Pareto principle.

Conflict Between Pareto Principle and All Non-welfarist Principles

Sen showed that one particular formulation of a libertarian principle, which carries the implication that externalities of a sort may not be regulated, violates the Pareto principle. Subsequently, it has been asked more broadly which notions of right, justice, and fairness conflict with the Pareto principle. The answer, it turns out, is that essentially all such notions do, as long as they do not depend exclusively on individuals' utilities – that is, unless they are a reformulation of welfarism.

To state the matter more precisely, we can contrast the individualistic SWF introduced previously, $W(U_1(x),\ldots,U_n(x))$, which by construction depends only on individuals' utilities, with the more generalized SWF, $Z(x)$, which also may be written as $Z(U_1(x),\ldots,U_n(x),x)$. Under the latter, social welfare may depend on anything and, in particular, need not depend exclusively on how the pertinent state x affects individuals' utilities. For example, notions of merit or desert concern whether certain actions or attributes are rewarded, principles of corrective or retributive justice demand that specific norm violations be followed by compensation or punishment, and so forth. Under each of these non-welfarist criteria, knowing each individual's utility in state x is insufficient information to form a social judgement.

Kaplow and Shavell (2001) prove that, if an SWF is not individualistic, then it violates the Pareto principle, if one makes a certain continuity assumption. The assumption is not that the SWF is continuous in all respects. (It is allowed, for example, that infinitesimal violation of some right might cause a discrete reduction in social welfare.) Rather, it is assumed that there exists some good that, if all individuals are given more of it, ceteris paribus (for example, holding rights violations constant), all will have a higher utility and, moreover, the value of the SWF changes continuously as the amount of that good is changed.

The proof is roughly as follows. First, if the SWF does not depend only on individuals' utilities, there must exist two states that are evaluated differently despite everyone's utilities being the same. That is, the non-welfarist SWF is supposed, in at least one instance, to rank states differently on account of a non-welfare difference. Now, taking whichever of the two states ranks lower, we can increase slightly everyone's allotment of the aforementioned good. By continuity, if that increase is sufficiently small, the lower-ranking state must still be ranked lower. However, since all individuals had equal levels of utility in the two initial states, every individual in the modified state now has greater utility, making it Pareto preferred despite the fact that the posited non-welfarist SWF ranks it lower. Hence, the Pareto principle is violated.

One way to understand the conflict between the Pareto principle and all non- welfarist principles is to reflect on the fact that a non-welfarist SWF by definition gives some weight in some instances to a factor independent of its effect on individuals' utilities. We can compare a state that is preferred on account of this non-utility factor to a state that is otherwise identical except that all individuals are slightly better off with respect to some commodity. In other words, a non-welfarist SWF, by its nature, sometimes sacrifices welfare, and nothing in logic rules out the possibility that the welfare sacrifice is borne pro rata.

Subsequent work has generalized and extended this theorem. Campbell and Kelly's (2002) survey notes that the proof in Kaplow and Shavell (2001) does not require the SWF to be a function rather than a binary relation; that this relation need not be fully transitive, only acyclic; and that only lower continuity is required. In a different vein, Suzumura (forthcoming) derives a sort of converse, namely, given Pareto indifference (if everyone is indifferent then society is indifferent – a principle implied by welfarism), social choice must respect the weak Pareto principle (the version defined at the outset of this entry) as well as the strong Pareto principle (if everyone weakly prefers one alternative and at least one individual strictly prefers it, then it is socially preferred). This theorem requires two additional assumptions: positive responsiveness of the social decision to individual preferences, and that, ceteris paribus, any utility level for an individual can be reached by adjusting the amount of a particular divisible good received by that individual.

Kaplow and Shavell (2002) also offer a complementary demonstration of the conflict between all non-welfarist principles and the Pareto principle. If one restricts attention to symmetric settings – those in which all individuals are identically situated – then any non-welfarist principle conflicts with the Pareto principle in every instance in which its ranking differs from a purely welfarist one. Because everyone is affected identically, it must be that, whenever any amount of aggregate welfare is sacrificed, each and every individual's welfare is sacrificed. The significance of this result is that many traditions favour assessing principles for guiding society in hypothetical situations that, because they are designed to create an impartial perspective, have a symmetric character. Consider, for example, the original position of Rawls (1971) – with important prior formulations thereof by Harsanyi (1953) and others – in which individuals are taken to have no knowledge of their own characteristics. Likewise, the injunctions of the Golden Rule and, relatedly, of Kant's (1785) categorical imperative demand, in essence, that one examine rules as if both positive and negative consequences were borne symmetrically by all. Since, as noted, all choices in symmetric settings involve strict Pareto rankings (except in cases in which all are indifferent), admitting a non-welfarist principle entails the view that the socially preferred state is systematically one in which everyone is worse off.

Perspectives on the Conflict

The Pareto criterion is a bedrock principle. Yet it conflicts with all non-welfarist principles – whether they pertain to rights, justice, or fairness – and some of these principles have apparent appeal. How may this tension be reconciled? That the Pareto principle should be seen as paramount is suggested by the rhetorical question: To whom is one doing right, providing justice, or being fair if every possible beneficiary is thereby made worse off? Additionally, as Sidgwick (1907) and others have queried, if something like utility does not underlie rights and related concepts, by what criterion is the proper list of rights determined in the first instance, and how in principle should the inevitable conflicts between different rights be resolved? A possible reconciliation is suggested by lines of thinking that trace their roots to prominent political economists of a prior era (among others), as more recently elaborated in Kaplow and Shavell (2002).

The relationship between the Pareto principle and other seemingly appealing principles can be understood by reference to what are known as two-level moral theories. (Act utilitarianism versus rule utilitarianism comes to mind, although that somewhat problematic distinction is subtly different from the one under consideration.) As suggested by Hume (1751), Mill (1861), and Sidgwick (1907), one can envision a first-level principle (such as utility) that provides our ideal assessment of states (corresponding to an SWF) and also numerous second-level principles (for example, that one should keep promises, tell the truth, not kill others) that are used as guides by individuals in their everyday conduct. Subsequent prominent statements of this view include Harrod (1936), Rawls (1955), and, most extensively, Hare (1981).

Put in a more explicit optimizing framework, the first-level principle serves as the objective function, and possible second-level principles constitute the universe of feasible policies. This feasible set is assumed to be constrained by limits of human nature and human institutions. Accordingly, the optimal scheme – taken here to consist of the optimal subset of second-level principles – will be only second best. The aforementioned limits render any attempt at direct implementation of the first-best criterion – commanding that everyone in their individual or institutional capacity act always so as to maximize social welfare – inferior to employment of second-best principles that, inevitably, deviate from the first-best criterion (welfare) in some instances. Two sets of rationales for this conception of the social maximization problem have been offered.

The first sort of justification is based on decision-making costs, complexity, limited information, limited self-control (for example, myopia), and so forth. Such considerations imply that all manner of behaviour, including some types that have no interpersonal effects, should be guided by rules. Moreover, given the nature of the problems that such rules are designed to address, it is inevitable that the rules will not require performance of a complete social welfare calculus and hence will sometimes command behaviour that differs from the first-best outcome. This conflict hardly makes the first-best principle any less of an ideal, just one that is not perfectly achievable in practice.

Second, the nature of human motivation, particularly the problem of cabining self-interest, provides another reason that sensible individual and institutional commands sometimes deviate from a pure concern for individuals' utilities, and thus offers another account of the conflict between the Pareto principle (viewed here as an aspect of the first-level social objective) and alluring non-welfarist principles (understood as second-level rules). Emphasized by Hume, Mill, and Sidgwick, and also by Smith (1790) and Darwin (1874), this strand of thinking is rooted in what may be called moral psychology. As a consequence of biological and social evolution, human emotions may help to channel behaviour in a positive fashion. Opportunism – whether through cheating, theft, or aggression – may be constrained by the prospect of guilt feelings or social disapprobation. Cooperation may be encouraged by anticipated positive internal sentiments or praise by others. Two familiar examples are the retributive urge, the prospect of which may deter aggression, and the desire for social approval, which may inhibit opportunism and encourage constructive collaboration. Given the limitations of biological evolution (limits on altruism as well as the tendency of evolved mechanisms to be specialized), constraints on social inculcation (including the fact that much is directed at young children), and the factors mentioned with regard to the first rationale for second-level rules, it is unsurprising that the resulting precepts sometimes deviate from the first best. Once again, this gap does not call into question the supremacy of the first-best ideal as a matter of principle. (Interestingly, however, this second explanation suggests that emotional force will be associated with moral criteria – various notions of what is right, just, or fair – that conflict with the Pareto principle, which helps explain why our intuitions may be in tension with pure welfarism in some settings.)

Both of these enduring strands of thought that help to reconcile the conflict between the Pareto principle and non-welfarist notions are related to the more recent upsurge of interest at the intersection of economics and psychology, often under the rubric of behavioural economics. Just as Tversky and Kahneman (1974) have stimulated research on heuristics and biases in a range of economic settings, Baron (1993) and others have documented similar phenomena – such as overgeneralization – in individuals' moral thinking. Likewise, many researchers, including Frank (1988) – following intervening provocative statements by Darwin (1874) and Wilson (1975) – have reinvigorated Smith's interest in human emotions as forces that guide human behaviour, although not always in an ideal manner.

The foregoing discussion suggests that, in regulating individuals' behaviour, various normative criteria that conflict with the Pareto principle may nevertheless usefully advance welfare and thus, at root, be consistent with the underlying force for that principle. These non-welfarist notions may also be relevant to the promotion of welfare for other, related reasons. As argued at length by Bentham (1822–23) in his constitutional writings and Mill (1859) in *On Liberty,* second-best rules obviously may play an important role in constraining government officials. In addition, since many of the non-welfarist criteria exist because of their relationship with the promotion of welfare, they may be useful proxy standards in some settings. Finally, due to the affective aspect of many non-welfarist principles, a complete welfarist account would incorporate them because they are in part constitutive of individuals' utilities. Note that, in each instance, because the relevance of non-welfarist criteria lies in the advancement of welfare, there is no conceptual inconsistency with the ultimate motivation for the Pareto principle even though the non-welfarist second- level rules on their face deviate from the posited first-level ideal.

In sum, a complete understanding of the relationship between the Pareto principle and other, possibly competing normative principles involves many dimensions. Formal analysis of these principles reveals the existence of an underlying, logical conflict. Examination of literatures in other fields of economics and in other disciplines, however, suggests a fundamental harmony.

See Also

- ▶ Arrow, Kenneth Joseph (Born 1921)
- ▶ Behavioural Economics and Game Theory
- ▶ Interpersonal Utility Comparisons
- ▶ Philosophy and Economics
- ▶ Sen, Amartya (Born 1933)
- ▶ Social Norms

Bibliography

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

Baron, J. 1993. *Morality and rational choice*. Boston: Kluwer Academic Publishers.

Bentham, J. 1822–23. *Securities against misrule and other constitutional writings for Tripoli and Greece*, ed. P. Schofield. Oxford: Oxford University Press, 1990.

Campbell, D., and J. Kelly. 2002. Impossibility theorems in the Arrovian framework. In *Handbook of social choice and welfare*, vol. 1, ed. K. Arrow, A. Sen, and K. Suzumura. Amsterdam: Elsevier Science.

Darwin, C. 1874. *The descent of man; and selection in relation to sex*, 2nd ed. Amherst: Prometheus Books, 1998.

Frank, R. 1988. *Passions within reason*. New York: W.W. Norton & Co.

Gibbard, A. 1974. A Pareto consistent libertarian claim. *Journal of Economic Theory* 7: 388–410.

Hare, R.M. 1981. *Moral thinking: Its levels, method, and point*. Oxford: Oxford University Press.

Harrod, R. 1936. Utilitarianism revised. *Mind* 45: 137–156.

Harsanyi, J. 1953. Cardinal utility in welfare economics and in the theory of risk-taking. *Journal of Political Economy* 61: 434–435.

Hume, D. 1751. *An enquiry concerning the principles of morals*, ed. T. Beauchamp. Oxford: Oxford University Press, 1998.

Kant, I. 1785. *Groundwork of the metaphysics of morals*, Trans. and ed. M. Gregor. Cambridge: Cambridge University Press, 1997.

Kaplow, L., and S. Shavell. 2001. Any non-welfarist method of policy assessment violates the Pareto principle. *Journal of Political Economy* 109: 281–286.

Kaplow, L., and S. Shavell. 2002. *Fairness versus welfare*. Cambridge, MA: Harvard University Press.

Mill, J.S. 1859. *On liberty*. London: J.W. Parker.

Mill, J.S. 1861. *Utilitarianism*, ed. R. Crisp. Oxford: Oxford University Press, 1998.

Pareto, V. 1906. *Manuale D'Economia Politica*. Milan: Società Editrice Libraria.

Rawls, J. 1955. Two concepts of rules. *Philosophical Review* 64: 3–32.

Rawls, J. 1971. *A theory of justice*. Cambridge, MA: Harvard University Press.

Robbins, L. 1935. *An essay on the nature and significance of economic science*, 2nd ed. London: Macmillan.

Robbins, L. 1938. Interpersonal comparisons of utility: A comment. *Economic Journal* 48: 635–641.

Sen, A. 1970. The impossibility of a Paretian liberal. *Journal of Political Economy* 78: 152–157.

Sen, A. 1992. Minimal liberty. *Economica* 59: 139–159.

Sidgwick, H. 1907. *The methods of ethics*, 7th ed. Indianapolis: Hackett Publishing Company, 1981.

Smith, A. 1790. *The theory of the moral sentiments*, 6th edn. Oxford: Oxford University Press, 1976.

Suzumura, K. Welfarism, individual rights, and procedural fairness. In *Handbook of social choice and welfare*, vol. 2, ed. K. Arrow, A. Sen, and K. Suzumura. Amsterdam: Elsevier Science (forthcoming).

Tversky, A., and D. Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *Science* 195: 1124–1131.

Wilson, E.O. 1975. *Sociobiology: The new synthesis*. Cambridge, MA: Harvard University Press.

Pareto, Vilfredo (1848–1923)

Alan Kirman

Abstract

Pareto made major contributions to a wide range of subjects covering mathematical economics, statistics, sociology and many others. In economics his name is mainly associated with general equilibrium, welfare economics and ordinal utility. Yet he insisted on the need to confront economic theories with empirical data as his work on income distribution shows. Furthermore, he was far from convinced of the rationality of individual economic behavior. Yet these aspects of his work have been put one side and he is now regarded essentially as the forerunner of the axiomatic school which reached its zenith in the Arrow–Debreu model. This is paradoxical for the latter has been shown to provide no empirically falsifiable propositions.

Keywords

Adaptive expectations; Arrow–Debreu model of general equilibrium; Barone, E.; Cardinal and ordinal utility; Central limit theorems; Compensation principle; Competitive equilibrium; Convexity; Economic cycles; Edgeworth box; Edgeworth, F.Y.; Equivalent surplus; Experimental methods in economics; Falsificationism; Fixed point theorems; Free trade; General equilibrium; Homo economicus; Hyperplanes; Imperfect competition; Income distribution; Inequality; Infant-industry protection; Integrability; Marginal utility of money; Marshall, A.; Mathematics and economics; Monopolistic competition; Multiple equilibria; New Welfare Economics; Non- tâtonnement process; Ophelimity; Ordinal utility; Pantaleoni, M.; Pareto efficiency; Pareto optimality; Pareto, V.; Pareto's law; Partial equilibrium; Perfect competition; Positive economics; Positivism; Preferences; Rationing; Social optimum; Social welfare function; Socialism; Surplus; Tâtonnement; Transitivity; Utility; Utility measurement; Walras, L.; Walras's Law

JEL Classifications

B31

Vilfredo Pareto's name is one of the most familiar in economics, with the universal use of 'Pareto optimality' and the Pareto distribution. Yet in 1968 Allais said, in his biography of Pareto, 'His influence on the development of economics as a science was felt only after considerable delay and has largely been confined to Italy and France'. There are several explanations for this neglect. First, as Chipman (1976) in his admirable survey of Pareto's contributions suggested, it has become less and less fashionable to cite early scholars. Second, Pareto spread his net wide and his persistent desire for empirical verification led him to explore the sociology of human behaviour. This part of his work had some success among sociologists, but interest in it has waned. Third, among theoretical economists Pareto's lasting contribution has come to be regarded as that which helped the evolution of economic theory on the path from the contributions of his predecessor Walras to the full axiomatic formulation of the Arrow–Debreu model. Yet this was but a small part of his overall contribution and of interest to an elite of theorists. Finally, Pareto has been wrongly described as the originator of some ideas and yet has not been credited with certain ideas which he did originate.

All of this explains why his reputation in economics was limited for a long time and has fluctuated considerably since his death in 1923.

Vilfredo Pareto was born in Paris in 1848 where his father, a follower of Mazzini, had exiled himself for political reasons and returned to Italy when Pareto was four years old. His family moved to Florence in 1862. In 1864 Vilfredo Pareto finished school at the early age of 16 and entered the University of Turin where he studied mathematics. He then went on to do a doctorate in engineering and his thesis was entitled, 'Principi fondamentali della teoria della elasticita dei corpi solidi e ricirche fondamentali sulla integrazione delle equazione diffenziali che ne differiscono l'equilibrio'. The fact that it was on the equilibrium of a physical system is, of course, significant for his later economic work.

Pareto then took up a post as an engineer for a railway company in Florence and continued in this work for three years. In Florence he became a member of the Accademia deo Georgofili and during this period he wrote a number of pieces in economics on such varied topics as the comparison of the advantages of publicly and privately owned railway systems, the merits of proportional representation and the state of the Italian industrial system. He was an ardent campaigner against any form of state interference with the market system and was one of the founders of the Adam Smith Society in Ferrara. As a consequence of their activity he developed a network of contacts in the fields of economics and politics. In 1880 and 1882 he was a candidate for member of parliament as an exponent of free trade, but was unsuccessful. In 1875 he was appointed as technical director of an ironworks in Florence. When raising the capital necessary to modernize the plant he travelled widely in Europe meeting bankers and financiers.

Through his intellectual activities, he established a relationship with Pantaleoni, who later held a chair at the University of Geneva and who introduced him to Walras in Switzerland in 1891. Pantaleoni then, in 1892, recommended him as a worthy successor to Walras at Lausanne when it became clear that the latter's health would not allow him to continue teaching. (See Walras 1965, vol. 2, p. 455, letter no. 1015.) He took up the chair in Lausanne in 1893 and his vision of the state of the field at that time is illustrated by his inaugural lecture in which he said.

> Nous ne connaissons la théorie d'aucun phénomène naturel dans tous ses details; nous connaissons seulement des théories des phénomènes ideaux, qui se rapprochent plus ou moins du phénomène concret.

After eight years at the university during which he produced his first major contributions to economics and was involved both in the activities of the university and of the canton de Vaud, he bought a villa near Geneva at Celigny. Here he progressively isolated himself from the university although, despite various health problems, he continued to teach in Lausanne till 1911. (Despite the fact that Schumpeter 1949, describes him as having had a 'vigorous and fertile old age', it is clear from his correspondence with the Dean of the Faculty at the university that Pareto was constantly preoccupied by health problems, in particular a heart ailment. See the letters from the archives of the law faculty cited by Biaudet 1975.)

The university organized a 'jubilee' in his honour in 1917 which was attended by a large number of distinguished social scientists and included an official delegation from the Italian government. The latter did something to offset Pareto's feeling that his own country had treated him badly.

While it is customarily asserted that Pareto did not start his work in economics until he reached the age of 45, he would obviously not have been offered a chair with no justification as to his ability in that subject. In fact, for well over ten years prior to his nomination in Lausanne Pareto had been interested in and had contributed to economics.

From the outset, Pareto was preoccupied by the idea of the economy as a complete system and by the interaction between the various sectors of the economy. In this, he was completely in line with the approach developed by Walras and far from the predominantly partial equilibrium analysis of his English contemporary, Marshall. What he was interested in was providing rigorous but parsimonious models of individual economic behaviour and then constructing from these a model of the economy as a whole. He was interested in the 'points of rest' of his system and their welfare characteristics.

It was his background as engineer and mathematician that led him to adopt a formal approach to the subject and his frustration with his inability to explain empirical facts that later led him to extend his analysis to sociology. This last phase of his career should not be construed as a disillusionment with the mathematical approach but rather as an attempt to include the other phenomena which he thought might account for the failure of economics to explain empirical facts. In particular he sought to include in his analysis the idea that people could make, from an economic point of view, 'irrational choices'. In so doing he anticipated the modern 'cognitive' approach to economics by a century. His overall aim was therefore to broaden his analysis and eventually to construct a system of laws capable of describing the behaviour of society as a whole, an enterprise that Schumpeter (1949) dismissed as a 'complete delusion'.

As has been observed his work received rather little attention for a long while. Allais (1968) lamented the failure of the economics profession to recognize the pioneering work of Pareto on the idea of social surplus. (This reflected, in part, his frustration with the lack of recognition of his own work in this area, which was later to be corrected by the award of the Nobel prize.) Georgescu-Rogen (1975) said, ‘There is no denying that Pareto’s own ideas met with an incredible lack of attention from most economists during his life as well as for many years after.’ Hicks (1932) wondered why economists had been so hesitant to study Pareto’s work and suggested that it was perhaps ‘the sheer impressiveness of his achievements’ that discouraged them. However, later Hicks (1975) himself, observed that the origins of Pareto’s contributions on the social optimum form could be traced to Edgeworth rather than to Pareto himself. Yet Malinvaud (1993) asserted that Pareto is now unanimously regarded as one of the founders of the Arrow–Debreu approach to theory. To see how this re-evaluation has taken place, we need to examine Pareto’s general approach to economics, certain of his specific contributions and his relationship to the work of his contemporaries and his predecessors.

It is worth observing at the outset that, while he condemned literary economists out of hand and professed to be interested only by a strictly scientific approach to economics, he, nevertheless, frequently made normative judgements and indulged in casual empiricism. This was, in part, a reflection of a world in which academics would feel much freer to express themselves on a wide variety of subjects without the many constraints that govern academic publication today.

Yet, despite a period in the desert, perhaps due to his having formed almost no students, some of Pareto’s main contributions have come to be recognized as having had a profound and lasting impact on economics. The three contributions to economics which have best stood the test of time are the *Cours d’économie politique* (1897), the *Manuale d’economia politica* (1906) and his article ‘Economie mathématique’ in *L’Encyclopédie des Sciences Mathématiques* (1911). In addition to these one has to mention his articles, in particular those collected and published later as *Marxisme et économie pure* in the *Oeuvres complètes* (1964–84) together with the *Trattato di sociologia generale* (1916) which, with *Les systèmes socialistes* (1901c) includes a substantial body of economic analysis.

The Cours

Of these contributions, the first, the *Cours*, originally published in two volumes, contains an exposition of economic theory illustrated with numerous empirical facts. The theory is presented in a more precise and refined way than that of his intellectual predecessor Walras and the emphasis, in the theoretical analysis is consistently and unequivocally on the interdependence of economic phenomena and the idea of general equilibrium. However, it should be noted that only 75 of the 800 pages are devoted to pure theory and that there is very little that is completely original. At least in Bousquet’s (1928a) eyes the theory was better presented than in Walras’s *Eléments* (1900) where he suggests the exposition is so tedious as to deter anyone from reading it. Nevertheless, the organization of the *Cours* is curious and, as Cirillo (1979) remarks, it is odd that production is treated after banking and social evolution. Of course, given its title and Pareto’s new responsibilities it is not surprising that it gives the strong impression of having been assembled from course notes, and that the order and content of these left something to be desired. (Pareto’s teaching cannot have been quite as discouraging as some have suggested since he wound up with 56 students in 1893 as opposed to the six who attended Walras’s last courses.) It is also somewhat odd that, given Pareto’s strong feelings about the importance of a positivist approach to economics he periodically indulges in direct pleading for the ‘liberal’ cause. It should, however, be remembered that, while so much of the material that Pareto discusses is now standard and has been refined by successive generations of economists, much of it was new, recent, or even original for him and it should be judged in context. What is remarkable is that Pareto, although one of the founders (if not *the* founder) of the school which culminated in the Arrow–Debreu model, did not hesitate to move beyond the strictly theoretical framework. He included empirical observations and examples of economic phenomena for which he was able to develop little satisfactory theory. Much of the statistical material in the *Cours* had, according to Pantaleoni (1924), been gathered while Pareto was a businessman, and this fact may have some bearing on its presentation. However, the *Cours* did also contain the first material on income distribution, a subject which will be examined below in more detail. This material provoked a great deal of comment and criticism from such authorities as Edgeworth (1925), and the latter was some 20 years later to comment ironically both on the universality of the law and on the character of its proponent.

The *Manuel*

The *Manuel* marks, as Ingrao and Israel (1990) suggest, a watershed between Pareto's involvement in economics and his move into sociology. (The reader seeking a detailed and rigorous account of Pareto's main theoretical contributions in the *Manuel* need look no further than Malinvaud 1993.) It illustrates the coexistence of philosophical reflection, empirical observations and rigorous analysis in Pareto's work. It explicitly acknowledges errors in the *Cours*, in particular that of taking too dogmatic a position in favour of free trade. While still accepting the theoretical arguments in favour of the free trade position, Pareto now had doubts as to the practical value of these arguments. He was no longer convinced that 'homo oeconomicus' was useful as anything other that a theoretical construction. He goes even further and even suggests that theoretical economics have not had, from a practical point of view, 'any great utility so far' (pp. vi–viii). This did not lead him, however, to abandon theory and he continued to develop his 'successive approximations' approach which he thought would lead from a highly abstract theory to a closer approximation of reality.

Of this book, it is the last section, the 'Mathematical Appendix' (in the French edition, 1909, pp. 538–671) which has come to be thought of as Pareto's basic contribution to the theory of general equilibrium and to what we now call 'Pareto optimality' but which he referred to as 'The maximum of society's ophelimity'. (This appendix was considerably modified and rewritten for the French edition, in large part as a result of Volterra's 1906, comments on the Italian edition.) Although the appendix with its formal analysis is the most widely cited part of the *Manuel*, it makes up less than a quarter of the volume. The rest of the work gives an insight into the more general view that Pareto had of economics.

The first two chapters give his views on the scientific status of the social sciences. He argued strongly that in economics and in the social sciences in general, there were underlying laws and structures which had to be determined, specified and tested by scientific methods. However, he, himself, was later to become more and more frustrated with the failure of economic theory to explain empirical facts. The third chapter provides an introduction to the idea of equilibrium which he now saw as a sort of balancing between 'tastes' and 'obstacles.' In this he had moved on from a more static concept as envisaged in the *Cours* and thought of a situation in which the 'obstacles' reacted to the 'tastes' and sought a resting point for these competing forces. (He does seem to have thought of his own approach, even in the *Cours*, vol. 1, p. 18, as being more dynamic than that of Walras, but it is difficult to see it as other than the solution of a static set of equations.) In Chapters 4 and 5, he then separates consumption and production and reduces the individual's problem, in each, to one of constrained optimization. Consumers follow paths of increasing utility until they are brought to a halt by the resistance of the 'obstacles'. Producers seek profit but face technological constraints. (These do not, in current terms, define a convex set. The introduction of fixed costs which cause this complication did, however, allow Pareto to reconcile zero profits and profit maximization.) In the sixth chapter, he then brings the markets together to talk of the general equilibrium of the system and to discuss its efficiency properties. Indeed the 'first theorem of welfare economics' makes its first clear appearance here. Two things are worth noting in passing. Pareto was well aware that the important thing for the individual was to maximize subject to the constraints that he perceives, which might or might not be the ones he faces in reality. Second, he systematically considered the possibility of monopolistic competition in parallel with that of perfect competition, although it is this treatment of the latter that is remembered today. In the case of imperfect competition the 'obstacles' change as the result of the individual's actions.

The chapters following those on economic equilibrium deal with a variety economic problems. Pareto deals with demographic problems and their consequences for the labour force but does not pursue the analysis of resultant effects on the labour market. He uses many factual illustrations, in dealing with these problems as well of those of natural resources, in particular land. He also treats capital and savings and the theory of the interest rate. The latter reveals an interesting discrepancy between Pareto's treatment of consumption and that of capital. When dealing with capital he clearly expressed the idea that goods are dated and that the rate of interest can therefore be deduced from the differences in successive prices. Yet when dealing with consumption the notion that the dates at which goods are available is part of their definition is much less clear. He then goes on to deal with monetary problems, but is clear that money can only be introduced once the general equilibrium problem has been fully analysed. All the latter problems are dealt with summarily and consist often of observations based on specific facts. The last chapter is devoted to 'concrete economic phenomena', although these already figure largely in the two preceding chapters.

Other Contributions by Pareto

There is little point in simply cataloguing Pareto's numerous other contributions but there are one or two specific items which cast light on the evolution of his thought. One aspect of his work that has been lost from sight is that on international trade, and yet his two articles on that subject had a clear influence on the major figures in the field. Ohlin (1924) went as far as to say that had he read Pareto earlier he would have saved himself a great deal of time and effort. Haberler (1965) said, 'But the only important theoretical advance has been the application, notably by Pareto, of general equilibrium analysis to the problems of international trade.' In particular, in marked contrast to the standard view that he was an unalloyed free trader, he used a sort of 'infant industry' argument in favour of protectionism for this, he thought, would lead to the emergence of a vigorous and productive class which would by its activities lead to a long-run gain which would more than offset the short-term loss from the absence of free trade. (This argument may be found in Chapter 9 of the *Manuel* and in the *Trattato.*) This, of course, was strongly related to his sociological theories, particularly that concerning the 'circulation des elites'.

A second and interesting aspect of Pareto's work was his concern with the origin and nature of economic cycles and crises. This does not fit well in a framework which is essentially static but he had the clear idea as early as the *Cours* that there was a certain overshooting in individuals' adaptation of their expectations. Thus he thought that people move too far from optimism to pessimism and that this leads to the sort of cyclical behaviour we observe in economies. In this he was arguing for a vision different from that of the 'rational expectations' school of today and rather more in line with the idea of 'adaptive expectations' but with coefficients which lead to over-adaptation.

Pareto examined socialism as a system for allocating resources in the *Systèmes socialistes* (1901c). Pareto saw socialism on the one hand as a threat to private property with its desire to extend the role of the state to the detriment of individual liberty but on the other as a force for change in society. He was, for example, not convinced by the redistributive goal of socialism. This was coloured by his own work on income distribution and the constancy that he thought he had found under many different institutional arrangements. The relation of this work to that of Barone is interesting. Pareto saw a role for a 'Ministry of Production' as a way of overcoming the difficulty of fixed costs in production. He envisaged a system of taxes on consumers to cover these costs, and goods would then be sold at cost price. This, he maintained, would restore efficiency despite the non-convexities present in the system. The shift in Pareto's ideas away from the purely liberal view towards a more subtle view of the role of the state is intimately linked to the evolution of his thoughts on sociology. He had what has been described as a 'clientelist' view of the organization of society. In his view, the mechanics of government intervention are governed by its need to satisfy its clients and not, as in the collectivist state view, to allocate resources efficiently and equitably.

This analysis together with Pareto's later emphasis on the non-rational, from an economic point of view, elements of choice led to his considering the achievement of any sort of economic efficiency, through a market system alone, as a utopian dream. He was, therefore, highly critical of those who insisted on applying economic theory without taking the whole political and social system into account.

From the economics point of view these contributions are not regarded as major and it is those that are regarded as of lasting importance that will be reviewed in the remainder of this contribution.

Ordinal Utility, Measurable Utility and the Integrability Problem

One of Pareto's major contributions has long been considered as that of establishing that an ordinal notion of utility is sufficient for the construction of equilibrium theory. The importance of this step for the development of modern theory was not immediately recognized by Pareto. It was only with his article (1900) that he started to develop a fully ordinal theory. Whether this led Pareto actually to reject the idea of 'measurable utility' is an interesting question. One suggestion is that he still adhered to the idea of some 'true measure' of utility but that he thought that is was simply impossible to identify the appropriate function. In fact, there is no logical contradiction between the observation that ordinality suffices to establish equilibrium and the idea that utility has some cardinal sense. Indeed, although he was very clear that any one of the 'indices' of utility would suffice for his analysis, he stated that 'In certain cases they would permit knowledge of the value of ophelimity' (*Manuel*, Mathematical Appendix).

Before returning to the measurability problem let us first examine how Pareto developed his ordinal approach. In the *Manuel*, he explicitly contrasts his analysis of 'indifference curves' which are constructed without reference to any utility function to that of Edgeworth who started with 'ophelimity' or 'utility' and obtained expressions for the indifference curves (*Manuel*, p. 540, n. 1).

Pareto proceeds as follows. In the case of two goods, consider x and y the quantities consumed of those goods and consider the quantities dx and dy which, when added to x and y, leave the consumers' satisfaction unchanged. This gives an equation:

$$f_1[x, y]\,dx + f_2[x, y]dy = 0.$$

This equation for the 'indifference line' gives us the expression: $dF[x, y] = 0$ which is satisfied by an infinite number of F, and Volterra (1906) remarked that 'Ophelimity is one of these functions F′.' (Whether this remark should be interpreted as meaning that any of these functions would serve as a utility function or whether the idea was that amongst these functions was the 'true utility function' is an interesting question.) In any event, as Volterra pointed out, Pareto's treatment required further work if it was to be extended to the case of three or more goods. The problem here is a simple one. The equations that Pareto wrote down define the tangent hyperplanes to the true indifference surfaces at each point in the consumption space, and what Volterra observed was that there may be no utility function compatible with these equations. This 'integrability problem' has preoccupied theorists until recently, although Allais (1973) dismisses it as a red herring.

Pareto's well-known (1906b) paper on 'Ophelimity in non-closed cycles' has generally been regarded as an attempt to solve this problem. There are many interpretations of this but one is that it revealed that convexity can replace transitivity. (However, Chipman 1971, suggests that Pareto's real aim was to give a full treatment of the measurable utility problem and Malinvaud 1993, regards it as a straightforward attempt to deal with the problem of the transitivity of preferences which in turn is directly linked to the problem of the existence of a utility function. Without entering into the details it is worth observing that Sonnenschein 1971, showed that the assumption of the convexity of preferences and hence of the quasiconcavity of the utility function could be substituted for transitivity. Thus transitivity of preferences is not a necessary condition for equilibrium analysis. Pareto himself, as Malinvaud 1993, indicates, glimpsed the importance of convexity and discusses it in Section 4 of the *Manuel* and in Sections 44–50 of the mathematical appendix.) Indeed convexity of preferences is, in a certain sense, a natural condition. This is simply because a great deal of theoretical work in economics boils down to looking at the solution of the maximization of a concave or quasi-concave function over a convex set, and for this most economists have settled for an examination of the firstorder conditions for such a maximum. Here the function in question is the utility function and the convex set is provided by the budget constraint or, in Pareto's terminology, the 'obstacles'. Given this, by now standard view, it is easy to understand why Pareto's 1906 contribution seems so convoluted and makes such difficult reading.

One of the problems is that there is, as in several of Pareto's works, a preoccupation with the order in which goods are consumed which, to modern eyes, confuses the discussion. In modern general equilibrium theory, all goods are dated and hence changing the order of consumption changes the specifications of the bundle of goods in question. Indeed, Pareto, unlike the other economists of his time, did, at some points in his analysis, explicitly adopt the idea of dating goods in order to reduce a temporal problem to a static one. But he also devotes considerable time to discussing 'paths of consumption'. (Detailed analysis of Pareto's analysis of the 'order of consumption' problem can be found in Chipman 1971, and Malinvaud 1993.) Thus, individuals move along a path improving their welfare until they arrive at the best bundle given their constraints. Was this how individuals actually behave as he originally indicated or did he, as in his 1906 paper, regard consumption paths as only involving time in an abstract and 'virtual' sense.

Pareto also considered explicitly in this context the measurability of utility. As it happens, as Chipman (1971) points out, in the specific case discussed by Pareto, utility is independent of the path of consumption. In this case, consider the marginal utility ('elementary ophelimity') of a good as dependent only on the quantity consumed of that good. (The type of result Pareto was aiming at is closely related to another case already treated by Fisher 1896.) The utility function obtained in this case is measurable, that is, it is invariant up to a linear transformation. Pareto's work here may thus best be regarded as an early attack on the problem of separable utility functions. Given this, it seems that, while Pareto recognized that equilibrium analysis could be carried out using only ordinal utility, he still had not reached the point of abandoning measurable utility as a concept, and still attached importance to the idea that the difference between the utility obtained in two situations had some significance. Thus, perhaps paradoxically to modern eyes, having liberated the theory of general equilibrium from the notion of cardinal utility Pareto continued to concern himself with the idea of measurable and comparable utilities.

The integrability problem mentioned above, that of recovering an underlying utility function from demand behaviour, was already solved in large part by Antonelli (1886) in a paper which Pareto had in his possession, albeit briefly, since he commented on it in a letter to Pantaleoni in 1891 (letter no. 39 in Pareto 1960). However, Pareto does not seem to have attached much importance to Antonelli's work, although Walras had already praised it. Thus, for some reason, Pareto did not profit from what had already been done and did not make any significant contribution in this particular direction despite assertions in the literature to the contrary. This may have been because of the brief acquaintance that he had with Antonelli's contribution, or may rather be, as Chipman maintains, because he was less concerned with this problem than with that of the measurability of utility.

To conclude the discussion of this part of his contributions, it is worth emphasizing that Pareto arrived at conditions for economic equilibrium using preferences alone and thus clearly marked out the trail for modern economic theory, but that he did not abandon his interest in the nature of utility and its measurement, and indeed it was as a result of this dual preoccupation that he adopted the curious term ‘ophelimity’.

Lastly, it is worth mentioning that Hutchison (1953) asserted that Pareto had anticipated Slutsky’s income and substitution effects. This would indeed have been a major achievement, but, as C. Weber (2002) has shown, a careful reading of the appropriate passage shows that this is not the case. This is yet another example of the inappropriate attributions involving Pareto’s work.

General Equilibrium

There can be few who have studied general equilibrium theory without using the famous Edgeworth box. This graphical trick which was for a period, unjustifiably called the Edgeworth–Bowley box, actually first appears in its modern form in Pareto (*Manuel*, 1906, p. 355). This was used by Pareto to motivate his ‘proofs’ of the welfare theorems in the general case. The box is the simplest case of what was Pareto’s constant preoccupation, namely, the economy as a complete system. He can be thought of as regarding the economy as one market rather than as many individual markets which could be studied, as in the Marshallian tradition, separately. He wrote down what are now considered to be the standard equilibrium conditions for the consumer side of economy, showing the equality of the marginal rate of substitution to the price ratio, normalizing the price of money, which he assumed to give direct utility, to one (*Manuel*, Mathematical Appendix). (Paradoxically, as Hildenbrand 1994, points out, these conditions are absent from the final culmination of Pareto’s theoretical work, the Arrow–Debreu model, and Debreu dismisses the use of calculus and confines himself until much later to convex sets and separating hyperplanes.)

Taking these equations:

$$U_x = \frac{U_y}{P_y} = \frac{U_z}{U_z} \dots$$

together with Walras’s Law:

$$(x - x_0) + P_y(y - y_0) + P(z - z_0) \dots = 0$$

and differentiating, he found expressions for

$$\frac{\partial y}{\partial P_y} \cdot \frac{\partial z}{\partial P_z} \text{ and so on.}$$

These he had already set out in his 1892 article in the *Giornale degli Economisti.* He then shows that if goods are independent, that is

$$U_{xy} = 0x\,\pi y,$$

then the conditions $U_{xx} < 0$ $U_{yy} < 0$ and so on. imply that demand for each good is a decreasing function of its own price (*Manuel*, Mathematical Appendix, section 53). This is a forerunner of more general but very recent results.

Pareto’s introduction of money into the utility function is in the spirit of his time and in so doing he was able to clarify some of Marshall’s analysis. Firstly, he showed in the *Manuel* (Mathematical Appendix, section 56) that, in general, the ‘marginal utility of money’ changes with prices. Thus it cannot arbitrarily be assumed to be constant.

Secondly, in *Cours*, section 83, he showed that the idea of estimating consumer surplus as the area under the consumer’s demand curve above the exchange price was wrong unless the marginal utility of money happened to be constant which, as we have just seen, it is not, in general.

Finally, he showed in his Encyclopaedia article (1911, section 23) that, if the elasticity of demand for all goods is constant, then it is unity, and remarked that, since Marshall had not realized that he had to impose this restrictive condition, his analysis was defective.

An important aspect of Pareto’s work that seems to have escaped attention is the effort that Pareto made to discuss what has come to be called ‘monopolistic competition’ and its introduction into equilibrium analysis (*Manuel*, chs. 3, 5 and 6). Recognizing that individuals can influence prices and that this should be taken into account led him to try to take explicit account of the demand with which they are faced. This merits two comments. Only recently has the introduction of 'monopolistic competition' into general equilibrium models resurfaced, and the article by Negishi (1961), which uses rather arbitrary assumptions, is generally cited as the first example. Thus Pareto was already dealing with a problem which has still not been really satisfactorily treated. The way in which the individuals have an effect on prices is revealed in his treatment of the 'obstacles' or constraints. As an individual modifies his demand he might be thought of as moving to the appropriate point on his budget hyperplane, thus as making a linear movement. But what if his displacement itself influences prices? In this case his movement will be nonlinear. Once again one can think of all these movements as being virtual and only the final result counting. Alternatively, one can think of a non-tâtonnement process in which individuals trade until they hit a constraint. In the Negishi style, non-tâtonnement process prices are still called centrally and then traders exchange and terminate before their desired bundle if they are not at the equilibrium prices. In Pareto's analysis trade is pairwise and there is rationing, leading Malinvaud (1993) to suggest that this is an early foretaste of the rationing literature associated with the names of Benassy, Drèze and Malinvaud. Nevertheless, it is clear that Pareto in this respect made a step in the direction of greater realism of the adjustment process.

A difficulty with all of Pareto's analysis is the confusion as to the nature of time. Indeed, he recognizes this himself, when talking of the passage from an initial position to an equilibrium, when he says (*Manuel*, ch. 3, section 171), that the issues he discusses fall into the domain of dynamics rather than statics. The modern convention that the adjustment process to equilibrium is instantaneous and that longterm dynamics consist of the passage from one equilibrium to another is far from that adopted by Pareto in this part of his analysis.

Pareto did not make a clear distinction between the question of existence and the question of stability. He regarded equilibrium as the terminating point of a process and this is brought out in the *Cours* and particularly in the *Manuel* (ch. 3, sections 110–15, for example). As has been suggested, the time taken for this process was not specified but is certainly not regarded, even conventionally, as negligible. Having described the passage or path from the initial position to the final position under assumptions of perfect competition with tâtonnement and pairwise non-tâtonnement processes, he then considered the monopolistic competition case, but rejected it as too difficult to handle. He did, however, enter into a discussion as to how individuals could push the economy towards a preferred equilibrium, from their point of view, in the case of multiple equilibria, and showed how they would try to manipulate the terms of exchange along this path (*Manuel*, p. 197) and discussed how individuals would benefit from doing this. Furthermore, in the light of this manipulation he suggested that certain equilibria would be stable. Thus Pareto recognized explicitly that stability is a property of a particular process.

Although he wished to consider equilibrium as the resting point of a process, Pareto did not try to show the existence of equilibrium as such except by counting equations and unknowns (*Manuel*, Appendix). In effect, he said that, since one could find the conditions for equilibrium, an interesting possibility would be to solve explicitly all the equations necessary to determine the equilibrium. However, as he pointed out (*Manuel*, ch. 3, section 217), 'If we take into account the fabulous number of equations that a population of forty million individuals and several thousand goods would give, it would not be mathematics that would come to the aid of economics, but economics that would come to the aid of mathematics' since such a system would be beyond human capacity to solve. Thus Pareto assumes from a simple argument the existence of a solution and simply dismisses the practicality of finding it for a large economy even if all the relevant equations were known. (This is in no way surprising since we now know that proving the existence of equilibrium is equivalent to proving the existence of a fixed point and the first fixed point theorem was proved by Brouwer in 1910.) (Pareto would have much appreciated Scarf's computational approach to the finding of equilibrium; Scarf and Hansen 1973.)

Thus the preoccupation with the formal establishment of equilibrium which was later to dominate mathematical economics was not shared by Pareto. Although relatively rigorous, he failed to specify various assumptions used for his approach, such as differentiability and only considered interior solutions to the maximum problem.

Before leaving Pareto's treatment of equilibrium, it is interesting to note that he was clearly aware of the possibility of multiple equilibria and in his diagrams in the *Manuel* (p. 192), he seems to have realized that 'in general, the number of equilibria would be odd', a result proved only very recently. An antecedent for this can be found in Edgeworth (1881), who explicitly talks about several equilibria and the fact that they will 'alternate' in terms of stability.

Pareto also suggests, as mentioned, that a collectivist state would be better able to lead its economy to an equilibrium than an economy based on private property. The reasoning given for this is based on Pareto's particular view of production and his introduction of non-convexities. This assertion, given Pareto's natural aversion to state intervention, is also heavily qualified (*Manuel*, ch. 6, sections 58–61). Nevertheless, it is interesting to note the contrast with the view of Pareto as an unqualified liberal.

Efficiency or 'Pareto Optimality'

Of all Pareto's contributions to economics, it is this notion of 'optimality' or efficiency that has made the greatest impact.

Yet it was not he who first gave a definition of a situation corresponding to the modern definition. Edgeworth (1881) clearly defined a situation in which the utility of each individual is maximized given the utilities of the others. Although this definition is given in the context of an exchange economy, its extension to more general cases was not difficult.

It was not so much the introduction of the idea but the use that Pareto made of it which makes his contribution important. Thus, although he had read Edgeworth, his definition, which also includes production, is an integral part of his own work.

Pareto defined a notion of surplus or gain, which is what is now referred to as 'equivalent surplus', this is, the amount of a given numeraire good which would leave the individual indifferent between his original bundle together with this quantity of numeraire and his original bundle together with some proposed change in all the commodities. At an optimum or efficient point there is no surplus. Put alternatively, one could think of the economy as moving along paths as individuals seek to profit from the surplus that exists until no further such change is possible. Thus an efficient situation is one in which no feasible change exists which would correspond to a positive surplus. The originality and correctness of Pareto's contribution has been questioned and Samuelson (1947) suggested that it was Barone (1908) who first dealt with this point correctly. In fact, as Allais (1973) points out, Barone acknowledged Pareto's priority and furthermore developed a less adequate, price-dependent, version of surplus.

The real insight that Pareto had was that this notion of efficiency or optimality was independent of all institutional arrangements and of all distributional considerations (*Cours*, vol. 2). Pareto then went on in the *Manuel* (ch. 6 and Mathematical Appendix, sections 145–52), to establish the 'first theorem of welfare economics', that a competitive equilibrium is a Pareto optimum and a tentative version of the 'second theorem', that any Pareto optimum can be obtained as a competitive equilibrium from an appropriate distribution of initial resources. The latter result is only suggested and is never clearly stated. Furthermore, both results are incomplete and even incorrect as a result of the confusion in the treatment of production. There are also a number of simple errors that creep into the exposition which confuse the argument. To take a simple example at two points in the *Manuel* (Appendix, sections 45 and 89) he says that some people will necessarily be better off and others worse off. Did he here envisage only movements along the efficient surface and therefore rule out changes which would make everyone worse off? This is not clear for at another point (*Manuel* (ch. 6, section 37) he clearly envisages everyone's welfare as declining. The key here is that he makes the correct statement in the case of a finite move but rules out the possibility of everybody being worse off in an infinitesimal move. It was Wicksell (1897) who pointed this out and Chipman (1976) mounted a rigorous defence of Pareto.

Pareto's ideas on the nature of efficiency evolved over time and in the *Trattato* (sections 2128–39), he showed that the maximization of any social welfare function W which was an increasing function of individual utility functions U_i

$$W = F(U_1, U_2, \ldots)$$

whether the U_i were defined over the consumption of all individuals or just restricted to individual consumption gave an optimum. Now as Pareto states (*Trattato*, pp. 1342–3), it is clear that in defining W a government would have to give weights to the different individuals. The idea of including the consumption of other individuals in the utility functions extends the scope of normal economic analysis to what were considered at the time and are still often thought of as 'sociological' considerations.

Pareto did not observe that by appropriately modifying F all optima could be generated. As Allais (1968) suggests, it is not clear that Pareto was fully aware of the impact of this contribution.

In addition to these contributions to welfare economics Pareto has been credited with the founding of the 'New Welfare Economics'. In particular, it is argued that in his 1894b article 'Il massimo di utilita dato dalla libera concorrenza' he introduced the Hicks–Kaldor compensation principle. However, as Kemp and Pezanis-Christou (1999) point out, Pareto argued that compensation should only be a consideration if it was actually carried out, and not just potentially possible. He spelled out the way in which it could be achieved by transfers between individuals, but did not go as far as saying that situation X is better than situation Y if transfers could be made from Y that would make everybody better off than in X.

Income Distribution, 'Pareto's Law'

One of Pareto's major contributions was to propose a 'law' governing the distribution of income. Here by distribution of income is meant the distribution of personal income amongst individual economic units and not the distribution of income between factors of production. The latter line was developed by Ricardo, in particular and was, of course, at the centre of the Marxian, neoclassical, and Keynesian debates. Pareto's interest and motivation were very different. On the one hand, it has often been suggested that his work in this area reflected the search for some sort of universal principles underlying economic behaviour. This would not explain why he chose this particular domain. The reason for his initial interest was rather his disagreement with the socialist proposals to undertake institutional reforms to make the distribution of personal income more equal. His initial work was published in an article in the *Gironale degli Economisti* in 1895 then in a memoire (1896b) on the 'income distribution curve'. Detailed discussion is given in the *Cours* (sections 957–65) and in the *Manuel* (ch. 7, sections 2–31).

In the course of analysing different data, Pareto was led to believe not only that he had established a functional form for income distributions which was essentially independent of institutional considerations but, even more remarkably, that the parameter of that function might well be the same across all countries and thus also independent of institutional arrangements. This would be enough to make any attempts to achieve a significant redistribution of income impossible. It is not surprising, given its social implications, that his contribution has been the source of controversy. Second, this work can be thought of as a pioneering piece of applied econometrics and not therefore as in keeping with the rather abstract image that is often painted of Pareto's work.

Three formulae were proposed by Pareto and the first and most widely cited of these is given by:

$$N(x) = \frac{A}{x^a}$$

where $N(x)$ is the number of people having an income greater than or equal to x. As has been frequently pointed out, it has obvious problems where either x tends to zero, or one increases x so that $N(x)$ goes to zero. Pareto's proposed a second form which mitigated these problems and which was to replace x by $x + q$ where q is a constant. Using his original form, Pareto then estimated values for *a* in particular for data for the UK collected by Giffen (reproduced in Giffen 1904). He obtained for 1843: a = 1.5 and for 1879/80: a = 1.35. Further computations for Prussia, Saxony, Paris and several Italian cities gave values around 1.5 with a maximum of 1.73. Pareto denied that his 'law' had the status of a physical law, and stated in an article in the *Journal of Political Economy* (1897b) that 'I should not be greatly surprised if some day, a well authenticated exception were discovered.' Nevertheless, he believed that the values of *a* that he found, *a* itself being a statistic, were sufficiently close for his law to be 'provisionally accepted as universal'. This statement is not wholly unambiguous since closeness of the estimated parameter values is not an indication that the functional form itself corresponds well to the data. Nevertheless, Pareto asserted that the values he obtained which he considered to be remarkably close, despite the different origins of the data, could not be attributed to chance.

Pareto was well aware that other functional forms might also fit the data well; for example, he estimated a distribution of the form:

$$N(x) = \frac{A}{(x+a)^{\alpha}} 10^{-\beta x}$$

where *a* and *b* like *a* are constants. He found a value of *b* so low that he concluded that a distribution of the second form that he had proposed

$$N(x) = \frac{A}{(x+a)^{\alpha}}$$

would suffice.

It is worth noting, in passing, that the three forms proposed by Pareto have a number of particular properties. The third form has finite moments for all r whereas this is only true for the first and second forms for r < *a*. When *a*. is less than or equal to 2 the first form has infinite variance and 'Pareto's law' is characterized by a fat right tail. In this case both the first and second forms belong to the Pareto-Lévy class of stable distributions. Indeed, Barbut (2000) has pointed out that the reason that distributions of the Pareto type occur so frequently was shown by Paul Lévy (1937). He showed that stable distributions other than the normal exhibit asymptotic behaviour of Pareto's second form with $0 < a < 2$. Hence, a central limit theorem of a certain type exists for heavy tailed distributions to which the standard central limit theorem does not apply. Pareto was thus credited with having removed the stranglehold of the Normal distribution!

It has been widely recognized since Pareto's time that other distributions provide more satisfactory fits for particular income data. Nevertheless, 'Pareto's law' gives empirically a satisfactory fit for the upper tail of the income distribution (the top 20 per cent according to Lydall 1968) but is clearly inconsistent with the lower end. This has resulted in a search for distributional forms which are close approximations of the Pareto form for the upper tail.

However, it is not the adequacy of Pareto's income distribution as a description of empirical data that has been controversial, it is rather the relation between 'Pareto's law' and the problem of income inequality that has been the subject of dispute. Pareto says that, if the number of individuals with an income over a certain level x in relation to the number of those below that level increases, then inequality diminishes (*Manuel*, ch. 7, section 24). Unfortunately, there was a printer's error in the *Cours*, and there the opposite is stated, although from the footnote (*Cours*, Livre III, section 965), it is clear what Pareto intended. There has since been considerable confusion about what Pareto actually said.

Let $N(h)$ be the number of individuals with income above h (the 'minimum income') and $N(x)$ the number above x with $x >$ h. Then, as Pareto says, if we define

$$U_x = \frac{N(x)}{N(h)}$$

then, 'income inequality will decrease as U_x increases' (*Cours*, Livre III, section 965; *Manuel*, p. 390, n. 2).

Allais (1968) interprets Pareto as saying the opposite, perhaps following the error in the *Cours*. Yet if we now proceed and assume that 'Pareto's law' holds, then we have:

$$U_x = \left(\frac{h}{x}\right)^{\alpha}.$$

Since $x > h$ by hypothesis, U_x decreases when a increases and income inequality increases. Allais makes an error in his argument and states that:

$$\frac{N(h)}{N(x)} = \left(\frac{h}{x}\right)^{\alpha},$$

an error identical to that made by Roy (1966). Since both Roy and Allais had started from the original mistake in the *Cours*, this further error should have led them to the same final conclusion as Pareto that income inequality varies in the same direction as a. Roy indeed arrives at this conclusion and contrasts it with the work of Gini and others. Allais made a further error and stated that Pareto believed that income inequality varied inversely with a. All this gives some indication of the sort of confusion that has surrounded Pareto's contribution. An explanation of these different interpretations can, however, be found and is that the authors mentioned were working with different basic hypotheses. If two distributions with the same mean income are compared, then the view that inequality increases with a is correct. If, on the other hand, one compares two distributions with the same minimum income, then Pareto's view is the appropriate one.

If a were a constant, then there would be little hope for policies aimed at reducing income inequality, as Pareto pointed out to those in favour of the socialist position. Lastly, Pareto's law has the peculiar feature that the ratio of the average income above x say $m(x)$ to x itself is a constant given by:

$$\frac{m(x)}{x} = \frac{\alpha}{\alpha - 1}.$$

Allais suggested that this might be taken as Pareto's index of inequality. If this were so, then it would decrease with a, the opposite of what Pareto intended.

Economics and Physics: Pareto's View

What is clear from both Pareto's analysis and that of many of his contemporaries such as Edgeworth, Jevons and Fisher is that they all shared a conviction that there was an analogy between economic systems and those of classical mechanics. Edgeworth (1881) was quite explicit in suggesting that a 'mécanique sociale' would take its place alongside the 'mécanique celeste'. Jevons (1905) said that economics resembles physics in that 'the equations employed do not differ in general character from those which are really treated in many branches of physical science'. Another contemporary, Cairnes (1875, the citations from Cairnes and Edgeworth are taken from Cohen, 1994) was even more explicit. He asserted that 'Political Economy is as well entitled to be considered a "positive science" as any of those physical sciences to which this name is commonly applied.' He went on to argue that the principles of economics have identical features to those 'of the physical principles which are deduced from the laws of gravitation and motion.'

The validity and consequences of such assertions have been examined at length by Mirowski (1989), Ingrao and Israel (1990) and Cohen (1994). The extent to which the analogy between physics and economics has ensnared economics in a position which it could have avoided had it found its source of inspiration elsewhere – for example, in biology, as Marshall suggested – is well documented by these authors.

Pareto himself made the remark that when examining the equations which have to be solved to determine an economic equilibrium someone well versed in mathematics or physics would say, 'These equations do not seem new to me, they are old friends. They are the equations of rational mechanics.'

He went so far, in the *Cours*, as to draw up a table of analogies between the two disciplines.

What is most interesting about this table is not the analogies themselves, which are, in some cases, inaccurate and misleading, but rather the caveats that are provided by Pareto. He seemed to be well aware, even at his early stage in his writings, of the dangers of taking the analogy too literally, and in this he distinguished himself from a number of his contemporaries. He understood that, when extended to the full social system, the physical analogy was highly tentative. Yet he had no other formal frame of reference within which to model the socioeconomic system. This led to his increasingly cautious attitude in using equations from physics, but did not deter him in his goal of modelling the whole social system rigorously. Given the reservations expressed by Pareto it seems unfair to lay the blame for the domination of classical mechanics as a mathematical framework for economics at his door. This only partially absolves him, for his attitude was essentially that physics provided an analogy but those parts of it that were inappropriate could be put to one side. As Mirowski (1989) points out, it is a common error to believe that all parts of the physics metaphor are equally dispensable. This misunderstanding has led, in part, to the persistence of the metaphor for it seems that we are free to weaken it as much as we wish till it is suitable for our purposes. Had this been recognized as erroneous economists might have strived harder for an alternative metaphor.

Economics and Its Relationship with the Other Social Sciences

Pareto's vision of the nature of the social sciences is reflected in his works on sociology (in particular the *Trattato*) and a certain number of his positions mark him out from his contemporaries and his successors. He developed and reinforced his idea that such sciences should be positive and went as far as criticizing his earlier work, taking the 'author' of the *Cours* to task for mixing ethical and positive considerations (*Manuel*, Preface). His defence of positivism was clearly associated with Comte's position (1830) and he was interested in developing a 'positive theory of economic policy'. He argued that 'laws' or relations deduced from specific assumptions should be tested empirically against 'observed statistical laws'. He went further, however, and unlike J.S. Mill (1844), who asserted that to verify hypotheses was not part of the business of science, a position supported by Friedman (1953) and Machlup (1955) and others, later argued that assumptions should be examined to see how reasonable they are (*Trattato*, section 59). The importance of Pareto's statistical work which reflected his standpoint has tended to be overlooked and has been dominated by analysis of his purely theoretical contributions. It cannot be repeated often enough that Pareto insisted on what he called the 'experimental method' as the *only* appropriate method appropriate for the social sciences and would not countenance wholly theoretical work which could not be empirically tested.

His approach to economics reflected a double position. Firstly, he shared Marshall's opinion that economic theory should be aimed at examining 'man as he is' and should not become an abstract intellectual exercise. Secondly, however, while he wished economics to be a relevant science, he condemned attempts to apply too readily economic theory to real problems. He believed that much harm had been done to the cause of 'scientific economics' by such hasty applications. This was, he thought, particularly dangerous since economic considerations could not be isolated from more general sociological concerns and to do so would lead to misleading and erroneous conclusions. His preoccupation with the analysis of nonrational behaviour adds force to this view.

Finally, it should be remembered that, while Pareto was with Weber among the first to expound the principles of 'positive social science', his view of the status of economics was ambiguous. He believed fundamentally, and in this he shared Comte's view, that there should be a universal scientific approach to social science. Yet he recognized the need for and desirability of specialized disciplines, although he regarded these as building blocks for a general approach. Thus while he was persuaded that certain aspects of economic phenomena are more quantifiable than many social phenomena, he was not prepared to isolate man's economic activity from his other functions. It should clearly be recognized that Pareto himself (Pareto 1917) considered his *Trattato* as his most important work and that he came to believe that the non-economic component of social phenomena dominates the economic part; and as he said, 'The most important error of the so-called liberal economist is not to recognise this.' Thus Pareto was progressively more convinced of the importance of the non-economic in explaining the evolution of society.

Conclusion

Pareto's economic contribution has acquired an increasing reputation over time, unlike his sociological work. Yet, as was suggested at the outset, it is disappointing that this reputation should be constructed on the basis of such a small part of his work. His strictly theoretical contributions are an essential part of modern general equilibrium theory. Yet here his education and training pushed him towards an equilibrium notion close to those of classical mechanics, and in a certain sense he helped to lock economics into an unhappily rigid framework.

Pareto's work covered a wide range of subjects. Within the field of economic theory, not only did he examine the nature and existence of a general equilibrium but he also considered what we would now refer to as the problem of 'imperfect competition', that is, the analysis of directly and consciously conflicting interests. Furthermore, his concern with statistical verification and his constant references to the idea that economic theories should be confronted with economic facts as in his examination of the problem of the form of income distributions are all central to an understanding of his contribution. He was preoccupied with the idea that economic theory fails to explain many phenomena, not because the theory itself is inadequate, but rather because that theory is just one part of a larger theoretical structure which should incorporate all social phenomena.

All of this illustrates the richness and diversity of Pareto's work. It is therefore paradoxical that, as Pareto's stature as one of the major figures in the development of economics has grown in recent years, most of this increased recognition has been based on a limited part of his most formal contributions. As has been observed, Pareto came to emphasize more and more the role of the non-economic in explaining social phenomena, yet he has come to be remembered essentially as the forerunner of the axiomatic school of economics where rationality is rigidly imposed. Perhaps the most ironic aspect of the evolution of Pareto's reputation is the current state of general equilibrium theory. While the most refined version of Pareto's theory, the Arrow–Debreu model, has been shown, thanks to the Sonnenschein–Mantel–Debreu results, to provide no empirically falsifiable propositions, Pareto himself was impatient with the idea of purely theoretical models which were not subject to falsification. How would he have reacted to the idea that his reputation was to be essentially based on his contribution to the construction of such a model? How far he was in spirit from the theory that he is claimed to have founded can be understood from a remark he made to a specialist in mathematical logic.

> I cannot admit that there is any rational method which is superior to the experimental method: I do not accept that one can study what should be; I, on the contrary, try to find out what exists in reality. (Pareto, 1964–84, vol. 19, 1027)

Selected Works

1892–93. Considerazioni sui principii fondamentali dell'ecconomia politica pura. *Giornale Degli Economisti* 4: 389–420, 485–512; 5: 119–157; 6: 1–37; 7: 279–321.

1893. La mortalità infantile e il costo dell'uomo adulto. *Giornale Degli Economisti* 7: 451–456.

1894a. Teoria matematica dei cambi forestieri. *Giornale Degli Economisti* 8: 142–173.

1894b. Il massimo di utilita dato dalla libera concorrenza. *Giornale Degli Economisti* 9: 48–66.

1895. Teoria del commercio internazional. *Giornale Degli Economisti* 9.

1896a. (A proposito di un libro del dottore Fornasari). Il modo di figurare i fenomeno economici. *Giornale Degli Economisti* 12: 75–87.

1896b. La curve della entrate e le osservazioni del professor Edgeworth. *Giornale Degli Economisti* 13: 439–448.

1896–7. *Cours d'économie Politique*, 2 vols. Lausanne: Librairie de l'Université. (Referred to as *Cours* in the text).

1897a. Aggiunta allo studio curva della entrate. *Giornale Degli Economisti* 14: 15–26.

1897b. The new theories of economics. *Journal of Political Economy* 5: 485–502.

1899. Quelques exemples d'application de la méthode de moindres carrés. *Journal de Statistique Suisse* 121–150.

1899. Tables pour faciliter l'application de la méthode de moindres carrés. *Journal de Statistique Suisse* 121–150.

1900. Sunto di alcuni capitoli di un nuovo trattato di economia pura del prof. Pareto. *Giornale Degli Economisti* 20: 216–235, 511–549. In Chipman et al. (1971).

1901a. Sul fenomeno economico. Lettera a Benedetto Croce. *Giornale Degli Economisti* 22: 131–138.

1901b. Le nuovo teorie economiche. Appunti. *Giornale Degli Economisti* 23: 235–252.

1901c. *Les Systèmes socialistes*, reprinted in Pareto (1964–84).

1902. Di un nuovo errore nello interpretare le teorie dell'economia matematica. *Giornale Degli Economisti* 25: 401–433.

1906. *Manuale d'economia Politica.* Milan: Societa Editrice Libraria. (Referred to as *Manuel* in the text). Revised and translated as Manuel d'économie politique, Paris: Giard et Brière, 1909. A new very complete and annotated edition of the *Manuale*, edited by A. Montesano, A. Zanni, and L. Bruni. Milan: Universita of Bocconi Press, was published in 2006.

1906b. L'ofelimità nei cicli non chiusi. *Giornale Degli Economisti* 33: 15–30. English translation in Chipman et al. (1971).

1907–8. L'interpolazione per la ricerca delle leggi economiche. *Giornale Degli Economisti* 34: 266–285; 36: 423–453.

1910. Walras. *Economic Journal* 20: 138–139.

1911. Economie mathématique. In *Encyclopédie des sciences mathématiques*, I (iv. 4). Paris: Teubner, Gauthier, Villars. Trans as 'Mathematical economics', International economic papers 5 (1955), 58–102. (Referred to as 'Economie mathématique' in the text.)

1913. Il massimo di utilità per una collettività in sociologia. *Giornale Degli Economisti* 46: 337–338.

1916. *Trattato di Sociologia Generale*, 4 vols. Florence: Barbera. (Referred to as *Trattato* in the text.) Trans. as *The mind and society*, ed. A. Livingston. New York: Harcourt Brace & Co., 1935.

1917. Pratica e Teoria. *L'Economista* 8(July): 540–542.

1918. Economia sperimentale. *Giornale Degli Economisti* 52: 1–18.

1960. *Lettere a Maffeo Pantaleoni 1890–1923*, 3 vols, ed. G. De Rosa. Roma: Banca Nazionale Del Lavoro.

All the above works together with a number of other articles, reviews and comments have been collected and published from 1964 on as Pareto, Vilfredo, Oeuvres complètes, 28 vols, ed. G. Busino. Geneva: Librairie Droz, 1964–84.

Bibliography

Accademia Nazionale dei Lincei. 1975. *Convegno Internazionale Vilfredo Pareto (Roma, 25–27 ottobre 1973)*. Roma: Accademia Nazionale dei Lincei.

Allais, M. 1968. Pareto, Vilfredo: Contributions to economics. In *Encyclopaedia of the social sciences*. New York: Macmillan.

Allais, M. 1973. Inequality and civilizations. *Social Science Quarterly*, Special issue for the 50th anniversary of Pareto's death, September.

Amoroso, L. 1938. Vilfredo Pareto. *Econometrica* 6: 1–21.

Antonelli, G.B. 1886. *Sulla Teoria Matematica della Economia Politica.* Pisa: Tipografia del Folchetto. Trans. J.S. Chipman and A.P. Kirman as '*On the mathematical theory of political economy*'. In Chipman et al. (1971).

Atkinson, A. 1975. *The economics of inequality.* Oxford: Oxford University Press.

Barbut, M. 2000. Pareto et la statistique. L'homme extrême de Pareto: sa postérité, son universalité. In *Pareto Aujourd'hui*, ed. B. Alain. Paris: PUF.

Barone, E. 1908. Il Ministro della produzione nello stato collettivista. *Giornale Degli Economisti* 37: 267–293, 391–414.

Benini, R. 1906. *Principi di statistica metodologica*. Torino: Unione Tipografica Editrice.

Biaudet, J-C. 1975. Vilfredo et Lausanne. In Accademia Nazionale dei Lincei (1975).

Borgatta, G. 1924. I rapporti fra la scienza economica e la sociologia nell'opera Paretiana. *Giornale degli Economisti* 39 (64): 81–89.

Borkenau, F. 1936. *Pareto*. London: Chapman & Hall.

Bousquet, G.H. 1927. *Introduction aux Systèmes Socialistes de Pareto*. Paris: Giard.

Bousquet, G.H. 1928a. *The work of Vilfredo Pareto*. Minneapolis: Sociological Press.

Bousquet, G.H. 1928b. *Vilfredo Pareto, sa vie et son oeuvre.* Paris: Payot. In *The invisible hand*, ed. B. Ingrao and G. Israel. Cambridge, MA: MIT Press, 1990.

Bousquet, G.H. 1963. Pareto et ses *Systèmes Socialistes. Cahiers de l'Institut de Science Appliquée* (Supplement no. 134): 25–32.

Busino, G. 1963. Pareto e le autorità di Losanna. *Giornale degli Economisti NS* 22: 260–303.

Cairnes, J.E. 1875. *The character and logical method of political economy.* New York: Harper & Brothers.

Cappa, A. 1924. *Vilfredo Pareto*. Turin: Gobetti.

Chipman, J.S. 1971. Introduction to Part II of Chipman et al. (1971).

Chipman, J.S. 1976. The Paretian heritage. *Revue européene des sciences sociales* 14 (37): 65–173.

Chipman, J.S., L. Hurwicz, M.K. Richter, and H.F. Sonnenschein (eds.). 1971. *Preferences, utility, and demand.* New York: Harcourt Brace, Jovanovich.

Cirillo, R. 1979. *The economics of Vilfredo Pareto*. London: Frank Cass and Co..

Cohen, I.B. 1994. *Interactions*. Cambridge, MA: MIT Press.

Comte, A. 1830. *Cours de Philosophie Positive*, 6 vols. Paris: Schleicher.

Davis, H.T. 1941. *The analysis of economic time series*. Bloomington: Principia Press.

Demaria, G. 1949. L'oeuvre économique de Vilfredo Pareto. *Revue d'Economie Politique* 59: 517–544.

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

Edgeworth, F.Y. 1896. Supplementary notes on statistics. *Journal of the Royal Statistical Society* 2: 533–534.

Edgeworth, F.Y. 1925. *Papers relating to political economy*, 3 vols. London: Macmillan for the Royal Economic Society.

Eisermann, G. 1961. *Vilfredo Pareto als Nationalökonom und Soziologe*. Tübingen: Mohr.

Fisher, I. 1896. (A review of) 'La courbe de la répartition de la richesse', by Vilfredo Pareto. *Yale Review* 5: 325–328.

Fossati, E. 1949. Pareto dans son et notre temps. *Revue d'Economie Politique* 59: 585–599.

Frechet, M. 1939. Sur les formules de répartition des revenus. *Revue de l'Institut International de Statistique* 7: 32–38.

Friedman, M. 1953. The methodology of positive economics. In *Essays in positive economics*. Chicago: University of Chicago Press.

Georgescu-Roegen, N. 1975. Vilfredo Pareto and his theory of ophelimity. In *Accademia Nazionale dei Lincei* (1975).

Giacolone-Monaco, T. 1957. *Pareto–Walras da un Carteggio Inedito (1891–1901)*. Padua: Cedam.

Gibrat, R. 1931. *Les Inégalités économiques*. Paris: Sirey.

Gide, C. 1917. Le jubilé Vilfredo Pareto. *Revue d'Economie Politique* 31: 426–433.

Giffen, R. 1904. *Economic inquiries and studies*. London: G. Bell and Sons.

Haberler, G. 1965. *The theory of international trade*. London: William Hodge.

Hicks, J.R. 1932. Marginal productivity and the Lausanne school. A reply. *Economica* 12: 297–300.

Hicks, J.R. 1975. Pareto and the economic optimum. In *Accademia Nazionale dei Lincei* (1975).

Hildenbrand, W. 1994. *Market demand: Theory and empirical evidence*. Princeton: Princeton University Press.

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

Ingrao, B., and G. Israel (eds.). 1990. The invisible hand. Cambridge, MA: MIT Press.

Jevons, W.S. 1905. *The principles of science: A treatise on logic and scientific method*. 2nd ed. London: Macmillan.

Johnson, N.O. 1937. The Pareto law. *The Review of Economics and Statistics* 19: 20–26.

Kemp, M.C., and P. Pezanis-Christou. 1999. Pareto's compensation principle. *Social Choice and Welfare* 16: 441–444.

Kirman, A. 1998. Vilfredo Pareto. In *Italian economists of the 20th century*, ed. F. Meacci. Cheltenham/Northampton: Edward Elgar.

Lévy, P. 1937. *Théorie de l'addition des Variables Aléatoires*. Paris: Gauthier-Villars.

Lydall, H.F. 1968. *The structure of earnings*. Oxford: Oxford University Press.

Machlup, F. 1955. The problem of verification in economics. *Southern Economic Journal* 22: 1–21.

Machlup, F. 1960. Operational concepts and mental constructs in model and theory formation. *Giornale degli economisti e annali di economia* 19: 553–582.

Machlup, F. 1964. Professor Samuelson on theory and realism. *American Economic Review* 54: 733–736.

Malinvaud, E. 1993. Le Manuel de Pareto et la Théorie Moderne des Prix. *Revue d'Economie Politique* 103: 157–189.

Mandelbrot, B. 1963. New methods in statistical economics. *Journal of Political Economy* 71: 421–440.

Marshall, A. 1920. *Principles of economics*. 8th ed. London: Macmillan. (first published in 1890).

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

Mirowski, P. 1989. *More heat than light*. Cambridge: Cambridge University Press.

Mortara, G. 1924. Pareto statistico. *Giornale degli Economisti* 64: 120–125.

Negishi, T. 1961. Monopolistic competition and general equilibrium. *Review of Economic Studies* 28: 196–201.

Ohlin, B.G. 1924. *Handelns Teori*. Stockholm: Centraltryckeriet.

Pantaleoni, M. 1907–1908. *Pure economics*. New York: Kelley and Macmillan, 1957. Originally published in Italian as ‘Elementi di economia pura’ and translated into English in 1898. *Giornale degli economisti* 34: 266–285; 36: 423–453.

Pantaleoni, M. 1923. Vilfredo Pareto. *Economic Journal* 33: 582–590.

Pantaleoni, M. 1924. In occasione della morte di Pareto reflessioni. *Giornale Degli Economisti* 44: 1–19.

Pirou, G. 1938. *Les Théories de l’équilibre économique: Walras et Pareto*. Rome: Giornale degli Economisti e Rivista de Statistica.

Roy, R. 1966. Preface to V. Pareto, *Statistique et économie mathématique.* Geneva: Librairie Droz.

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

Samuelson, P.A. 1963. Problems of methodology: Discussion. *American Economic Review: Papers and Proceedings* 53: 231–236.

Samuelson, P.A. 1964. Theory and realism: A reply. *American Economic Review* 54: 736–739.

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

Schumpeter, J. 1942. *Capitalism, socialism and democracy.* New York: Harper & Brothers.

Schumpeter, J. 1949. Vilfredo Pareto (1848–1923). *Quarterly Journal of Economics* 63: 147–173.

Schumpeter, J. 1954. *History of economic analysis*. New York: Oxford University Press.

Sonnenschein, H. 1971. Demand theory without transitive preferences: With applications to the theory of competitive equilibrium. In Chipman et al. (1971).

Spengler, J.J. 1944. Pareto on population 1. *Quarterly Journal of Economics* 58: 593–598.

Stark, W. 1963. In search of the true Pareto. *British Journal of Sociology* 14: 103–112.

Volterra, V. 1906. L’Economia matematica e il Nuovo manuale del Prof. Pareto. *Giornale degli economisti* 32: 296–301. Translated by A.P. De Kirman, revised and edited by J.S. Chipman as ‘*Mathematical economics and professor Pareto’s New Manual*’, in Chipman et al. (1971).

Walras, L. 1900. *Eléments d’économie politique pure ou théorie de la richesse sociale*, 4th edn. Paris: Rouge, Lausanne, and Pichon. Last revised edition, Paris, 1926; English translation, *Elements of pure economics*, ed. W. Jaffé. London: Allen & Unwin, 1954. In *The invisible hand*, ed. B. Ingrao and G. Israel (1990), Cambridge, MA: MIT Press.

Walras, L. 1965. *Correspondence of Léon Walras and related papers*, 3 vols, ed. W. Jaffé. Amsterdam: North-Holland.

Weber, M. 1913. *Wirtschaft und Gesellschaft.* Originally printed as the third part of *Grundriss der Sozialökonomik*, 2 vols. Tübingen: Mohr. Reprinted in *Gesammelte Aufsätze zur Wissenschaftslehre*, Tübingen: Mohr, 1922.

Weber, M. 1949. *The methodology of the social sciences.* (trans: E. Shils and H. Finch). Glencoe: Free Press.

Weber, C. 2002. Did Pareto discover income and substitution effects? On an interpretation suggested by Hutchison. *Economics Bulletin* 2 (2): 1–6.

Wicksell, K. 1897. Vilfredo Pareto, Cours d’économie politique. *Zeitschrift für Volkswirtschaft, Sozialpolitik und Verwaltung*, 159–166. Reprinted in English in *Selected Papers of Knut Wicksell*, ed. E. Lindahl. Cambridge, MA: Harvard University Press, 1958.

Parnell, Henry Brooke (1776–1842)

Barry Gordon

A significant figure in the history of economic policy, Parnell was an Irish landowner, baronet (1812) and first Baron Congleton (1841). He was associated with the liberal wing of the Whig party in the British parliament. During the period that David Ricardo was in the Commons (1819–23), he lent strong support to Ricardo’s advocacy on economic policy issues. After Ricardo’s death, Parnell continued to press for tariff reform, repeal of the Corn Laws, and greater economy in government expenditure.

Parnell entered the Commons in 1802. Subsequently, he was a member of the Bullion Committee (1810), chairman of the Committee on the Corn Trade (1813), and chairman of the Committee on Public Income and Expenditure (the Finance Committee) from 1828. His motion of 1830 concerning the Civil List precipitated the resignation of the Duke of Wellington as Prime Minister. In 1831, Parnell was appointed Secretary at War, but was dismissed in 1832. Under Lord Melbourne, he became Treasurer of the Navy and Paymaster-General in 1835. He retained these posts until his death in 1842. For the last 20 years of his life he was an active member of the Political Economy Club.

A diligent committee man, Parnell also intervened frequently in economic debate on the floor of the Commons. Those interventions demonstrate considerable analytical ability as well as unswerving allegiance to the doctrines of economic liberalism. Particularly notable is his speech on the Customs Consolidation Bill (1825) in which he spells out what is required for a rational, utilitarian tariff reform according to Ricardian principles.

Publications by Parnell include several on currency and banking (1804, 1805, 1827 and 1832), and on the corn laws (1809 and 1814). His most influential work, however, was *On Financial Reform* (1830; 4th edn, 1832). This publication is a complement to his parliamentary efforts as chairman of the Finance Committee. Parnell was zealous for both the utmost economy in public spending and restructuring of the taxation system. Such restructuring, he believed, should include the removal of taxes on raw materials and semi-finished goods, general reductions in import duties, and the introduction of an income tax of between one-and-a-half and two per cent. These policy measures look forward to those of Sir Robert Peel in the 1840s and, even further, to those of Gladstone.

Selected Works

1804. *Observations upon the state of the currency of Ireland*, 3rd ed, with additional Appendix, 1804.

1805. *The principles of currency and exchange*.

1809. *Treatise on the corn trade and agriculture*.

1814. *The substance of the speeches of Sir Henry Parnell, bart., in the House of Commons, with additional observations on the corn laws*.

1827. *Observations on paper money, banking and over-trading*. London: James Ridgway. 2nd ed, 1829.

1830. *On financial reform*, 4th ed, enlarged, 1832.

1832. *A plain statement of the power of the bank of England*.

1833. *A treatise on roads*, 2nd ed, enlarged, 1838.

References

Fetter, F.W. 1980. *The economist in Parliament: 1780–1868*. Durham: Duke University Press.

Gordon, B. 1979. *Economic doctrine and Tory liberalism, 1824–1830*. London: Macmillan.

Hilton, B. 1977. *Corn, cash, commerce: The economic policies of the Tory governments, 1815–1830*. Oxford: Oxford University Press.

Parsons, Talcott (1902–1979)

J. Goodwin

Talcott Parsons was perhaps the most ambitious and influential sociologist of his generation. Parsons was born at Colorado Springs, Colorado, on 13 December 1902. He was educated at Amherst College (Massachusetts), the London School of Economics, and Heidelberg University, from which he received a doctorate in economics in 1927. Parsons served on the faculty at Harvard University from 1927 until his retirement in 1973. He played a key role in the organization of the interdisciplinary Department of Social Relations at Harvard (now defunct), serving as chair of that department from 1946 to 1956. Parsons was a prolific, if notoriously abstruse writer, producing more than a dozen major books and scores of articles on a variety of (mainly theoretical) subjects. His most influential works are *The Structure of Social Action* (1937) and *The Social System* (1951). Parsons died in Munich on 8 May 1979.

While at Heidelberg, Parsons came under the influence of Continental sociology, particularly the work of Max Weber and Emile Durkheim. Parsons was subsequently responsible, more than any other figure, for introducing the thought of these theorists into Anglo-American sociology. He translated Weber's famous essay on *The Protestant Ethic and the Spirit of Capitalism* into English in 1930; he analysed the work of Durkheim and Weber at length in *The Structure of Social Action* (1937); and his edited version of Weber's *Economy and Society* appeared in 1947 under the title *The Theory of Social and Economic Organization*.

Parsons himself attempted to elaborate nothing less than a comprehensive theory of society, a general theory of 'the social system'. He argued that the economy is a functional subsystem of the larger social system and that economic theory, consequently, is a 'special case' of the general theory of society. Parsons put forth the controversial claim that the thought of Marshall, Pareto, Durkheim, and Weber 'converged' on what he called a 'voluntaristic theory of action'. This theory emphasizes the normative and purposive dimensions of social (including economic) behaviour, rejecting purely utilitarian or interest-based accounts. Parsons also emphasized the importance of shared cultural values in his treatment of the 'Hobbesian problem' of social order and in his analyses of social change.

Parson's 'grand theory' continues to generate considerable debate. For his proponents, Parson's work ranks among the most sophisticated attempts to overcome the antinomies of modern social thought; for his critics, Parsons's ponderous prose conceals a simplistic and fundamentally conservative cultural determinism.

Selected Works

1937. *The structure of social action.* New York: McGraw-Hill.

1951. *The social system.* New York: Free Press.

Partial Identification in Econometrics

Charles F. Manski

Abstract

Econometricians long thought of identification as a binary event: a parameter is either identified or not. Empirical researchers combined available data with assumptions that yield point identification, and reported point estimates of parameters. Yet there is enormous scope for fruitful inference using weaker and more credible assumptions that partially identify parameters. Until recently, study of partial identification was rare and fragmented. However, a coherent body of research took shape in the 1990s and has grown rapidly. This research has yielded new approaches to inference with missing outcome data, analysis of treatment response, and other important problems of empirical research.

Keywords

Counterfactuals; Discrete response analysis; Errors in variables; Gini coefficient; Identification region; Law of Total Probability; Nonparametric methods; Parametric prediction; Partial identification in econometrics; Reverse regression; Sampling; Statistical inference; Treatment response

JEL Classifications

C14

Suppose that one wants to use sample data to draw conclusions about a population of interest. Econometricians have long found it useful to separately study identification problems and problems of statistical inference. Studies of identification characterize the conclusions that could be drawn if one were able to observe an unlimited number of realizations of the sampling process. Studies of statistical inference characterize the generally weaker conclusions that can be drawn given a sample of positive but finite size. Koopmans (1949, p. 132) put it this way in the article that introduced the term 'identification':

> In our discussion we have used the phrase 'a parameter that can be determined from a sufficient number of observations.' We shall now define this concept more sharply, and give it the name *identifiability* of a parameter. Instead of reasoning, as before, from 'a sufficiently large number of observations' we shall base our discussion on a hypothetical knowledge of the probability distribution of the observations, as defined more fully below. It is clear that exact knowledge of this probability distribution cannot be derived from any finite number of observations. Such knowledge is the limit approachable but not attainable by extended observation. By hypothesizing nevertheless the full availability of such knowledge, we obtain a clear separation between problems of statistical inference arising from the variability of finite samples, and problems of identification in which we explore the limits to which inference even from an infinite number of observations is suspect.

For most of the 20th century, econometricians commonly thought of identification as a binary event – a parameter is either identified or it is not.

Empirical researchers applying econometric methods combined available data with assumptions that yield point identification and they reported point estimates of parameters. Many economists recognized with discomfort that point identification often requires strong assumptions that are difficult to motivate. However, they saw no other way to perform inference.

Yet there is enormous scope for fruitful inference using weaker and more credible assumptions that partially identify population parameters. A parameter is partially identified if the sampling process and maintained assumptions reveal that the parameter lies in a set, its 'identification region', that is smaller than the logical range of the parameter but larger than a single point. Estimates of partially identified parameters generically are set-valued; a natural estimate of an identification region is its sample analog.

Until recently, study of partial identification was rare and fragmented. Frisch (1934) and Reiersol (1941) developed sharp bounds on the slope parameter of a linear regression with errors-in-variables, with refinement by Klepper and Leamer (1984) and others. Duncan and Davis (1953) used a numerical example to show that the ecological inference problem of political science is a matter of partial identification. Cochran et al. (1954) suggested conservative analysis of surveys with missing data due to non-response by sample members, although Cochran (1977) subsequently downplayed the idea. Peterson (1976) initiated study of partial identification of the competing risks model of survival analysis.

For whatever reason, these scattered contributions remained at the fringes of econometric consciousness and did not spawn systematic study of partial identification. However, a coherent body of research took shape in the 1990s and has grown rapidly. The new literature on partial identification emerged out of concern with traditional approaches to inference with missing outcome data. Empirical researchers have commonly assumed that missingness is random, in the sense that the observability of an outcome is statistically independent of its value. Yet this and other point-identifying assumptions have regularly been criticized as implausible. So it was natural to ask what random sampling with partial observability of outcomes reveals about outcome distributions if nothing is known about the missingness process or if assumptions weak enough to be widely credible are imposed. This question was posed and partially answered in Manski (1989), with subsequent development in Manski (1994, 2003, chs.1 and 2), Scharfstein et al. (2004), Blundell et al. (2004) and Stoye (2005).

Study of inference with missing outcome data led naturally to consideration of conditional prediction and analysis of treatment response. A common objective of empirical research is to predict an outcome conditional on given covariates, using data from a random sample of the population. Often, sample realizations of outcomes and/or covariates are missing. Horowitz and Manski (1998, 2000) and Zaffalon (2002) study nonparametric prediction when nothing is known about the missingness process; Horowitz et al. (2003) and Horowitz and Manski (2006) consider the computationally challenging problem of parametric prediction. Missing data on outcomes and covariates is the extreme case of interval measurement of these variables. Manski and Tamer (2002) study conditional prediction with interval data on outcomes or covariates, while Haile and Tamer (2003) analyse an interesting problem of interval data that arises in econometric analysis of auctions.

Analysis of treatment response must contend with the fundamental problem that counterfactual outcomes are not observable; hence, findings on partial identification with missing outcome data are directly applicable. Yet analysis of treatment response poses much more than a generic missing-data problem. One reason is that observations of realized outcomes, when combined with suitable assumptions, can provide information about counterfactual ones. Another is that practical problems of treatment choice as well as other concerns motivate research on treatment response and thereby determine what population parameters are of interest. For these reasons, it has been productive to study partial identification of treatment response as a subject in its own right. This stream of research was initiated independently in Robins (1989) and Manski (1990).

Subsequent contributions include Manski (1995, 1997a, 1997b), Balke and Pearl (1997), Heckman et al. (1997), Hotz et al. (1997), Manski and Nagin (1998), Manski and Pepper (2000), Moinari (2002), and Pepper (2003). The normative problem of treatment choice when treatment response is partially identified is studied in Manski (2000, 2002, 2005a, 2005b, 2006) and Brock (2005).

Another broad subject of study has been inference on the components of finite probability mixtures. The mathematical problem of decomposition of mixtures arises in many substantively distinct settings, including contaminated sampling, ecological inference, and conditional prediction with missing or misclassified covariate data. Findings on partial identification of mixtures have application to all of these subjects and more. Research on this subject includes Horowitz and Manski (1995), Bollinger (1996), Cross and Manski (2002), Dominitz and Sherman (2004), Kreider and Pepper (2004), and Molinari (2004).

There has been other research as well. In discrete response analysis, response-based sampling poses a 'reverse regression' problem in which one seeks to learn the distribution of outcomes given covariates but the sampling process reveals the distribution of covariates given outcomes. This problem has been studied in Manski (1995, ch. 4; 2001, 2003, ch. 6) and King and Zeng (2002). In econometric analysis of multi-player games, a long-standing problem has been to infer behaviour from outcome data when the game being studied may have multiple equilibria. Ciliberto and Tamer (2004) address this problem.

Whatever the specific subject under study, a common theme runs through the new literature on partial identification. One first asks what the sampling process alone reveals about the population of interest and then studies the identifying power of assumptions that aim to be credible in practice. This conservative approach to inference makes clear the conclusions one can draw in empirical research without imposing untenable assumptions. It establishes a domain of consensus among researchers who may hold disparate beliefs about what assumptions are appropriate.

It also makes plain the limitations of the available data. When credible identification regions turn out to be large, researchers should face up to the fact that the available data do not support inferences as tight as they might like to achieve.

The remainder of this article uses the problem of inference with missing outcome data and the analysis of treatment response to develop the common theme of recent research on partial identification and to give illustrative findings. Readers who aim to learn more may want to begin with two monographs that provide self-contained expositions with different audiences in mind. Manski (1995) presents basic ideas in a way intended to be broadly accessible to students and researchers in the social sciences. Manski (2003) develops the subject in a rigorous manner meant to provide the foundation for further study by econometricians.

Readers who prefer to learn about econometric methods through the study of empirical applications will find diverse case studies using observational data to analyse treatment response. Manski et al. (1992) investigate the effect of family structure on children's outcomes, and Hotz et al. (1997) analyse the effect of teenage childbearing. Manski and Nagin (1998) study the effects of judicial sentencing on criminal recidivism. Pepper (2000) examines the intergenerational effects of welfare receipt. Manski and Pepper (2000) and Ginther (2002) analyse the returns to schooling.

There have also been empirical studies of problems of partial identification that arise in analysis of randomized experiments. Horowitz and Manski (2000) study a medical clinical trial with missing data on outcomes and covariates. Pepper (2003) asks what welfare-to-work experiments reveal about the operation of welfare policy when case workers have discretion in treatment assignment. Scharfstein et al. (2004) analyse an educational experiment with randomized assignment to treatment but non-random attrition of subjects.

Inference with Missing Outcome Data

To formalize the missing data problem, let each member j of a population J have an outcome y_j in a space Y. The population is a probability space and y: $J \rightarrow Y$ is a random variable with distribution

$P(y)$. Let a sampling process draw persons at random from J. However, not all realizations of y are observable. Let the realization of a binary random variable z indicate observability; y is observable if $z = 1$ and not observable if $z = 0$.

By the Law of Total Probability.

$$\begin{aligned} P(y) = P(y|\, z = 1)P(z = 1) \\ + P(y|\, z = 0)P(z = 0). \end{aligned} \tag{1}$$

The sampling process reveals $P(y|z = 1)$ and $P(z)$, but is uninformative regarding $P(y|z = 0)$. Hence, the sampling process partially identifies P (y). In particular, it reveals that $P(y)$ lies in the identification region

$$\begin{aligned} &H[P(y)] \\ &= [P(y|\, z = 1)P(z = 1) + \gamma P(z = 0), \gamma \in \Gamma_Y], \end{aligned} \tag{2}$$

where Γ_Y is the space of all probability distributions on Y.

The size of the identification region $H[P(y)]$ grows with $P(z = 0)$, which measures the prevalence of missing data. The region is a proper subset of Γ_Y whenever the probability of missing data is less than 1, and it is a singleton when there are no missing data. Thus, $P(y)$ is partially identified when $0 < P(z = 0) <1$ and is point-identified when $P(z = 0)= 0$.

Means of Bounded Functions of y

A common objective of empirical research is to infer parameters of a probability distribution. The identification region for a parameter of $P(y)$ follows immediately from $H[P(y)]$. Let $\tau(\cdot):\Gamma_Y \rightarrow T$ map probability distributions on Y into a parameter space T and consider inference on the parameter $\tau[P(y)]$. The identification region consists of all possible values of the parameter. Thus,

$$H\{\tau(P(y))\} = \{\tau(\eta), \eta \in H[P(y)]\}. \tag{3}$$

Result (3) is simple but is too abstract to be useful as stated. Research on partial identification has sought to characterize $H\{\tau\ [P(y)]\}$ for different parameters. Manski (1989) does this for means of bounded functions of y, Manski (1994) for quantiles, and Manski (2003, ch. 1) for all parameters that respect first-order stochastic dominance. Blundell et al. (2004) and Stoye (2005) characterize the identification regions for spread parameters such as the variance, inter-quartile range, and the Gini coefficient; these authors apply their findings in empirical research assessing nationwide income inequality using surveys with missing income data.

The results for means of bounded functions are easy to derive and instructive, so I focus on these parameters here. Let R be the real line. Let $g(\cdot)$ be a function that maps Y into R and that attains finite lower and upper bounds $g_0 = \min_{y \in Y} g(y)$ and $g_1 = \max_{y \in Y} g(y)$. The problem of interest is to infer $E[g\ (y)]$.

The Law of Iterated Expectations gives

$$\begin{aligned} E[g(y)] &= E[g(y)|\, z = 1]P(z = 1) \\ &+ E[g(y)|\, z = 0]P(z = 0). \end{aligned} \tag{4}$$

The sampling process reveals $E[g(y)|z = 1$ and $P(z)$, but is uninformative regarding $E[g(y)|$ $z = 0$, which can take any value in the interval $[g_0,\ g_1]$. Hence, the identification region for E $[g(y)]$ is the closed interval

$$\begin{aligned} H\{E[g(y)]\} = [E[g(y)|\, z=1]P(z=1) \\ +g_0 P(z=0), E[g(y)|\, z=1]P(z=1) \\ +g_1 P(z=0)]. \end{aligned} \tag{5}$$

$H\{E[g(y)]\}$ is a proper subset of $[g_0, g_1]$ whenever $P(z = 0)$ is less than one. The width of the region is $(g_1 - g_0)P(z = 0)$. Thus, the severity of the identification problem varies directly with the prevalence of missing data.

Result (5) has many applications. Perhaps the most far-reaching is the identification region it implies for the probability that y lies in any non-empty, proper set $B \subset Y$ Let $g_B\ (\cdot)$ be the indicator function $gB(y) \equiv 1\ [y \in B]$; that is, $g_B(y) = 1$ if $y \in B$ and $g_B(y) = 0$ otherwise. Then $g_B(\cdot)$ attains its lower and upper bounds on Y, these being 0 and 1. Moreover, $E[g_B(y)] = P(y \in B)$ and $E[g_B(y)|z = 1] = P(y \in B|\ z =1)$. Hence,

$$H[P(y \in B)] = [P(y \in B|\, z = 1)P(z = 1),\\ P(y \in B|\, z = 1)P(z = 1) + P(z = 0)]. \qquad (6)$$

Observe that the width $P(z = 0)$ of this interval depends only on the prevalence of missing data, not on the form of set B.

When y is real-valued, result (6) immediately yields the identification region for the distribution function of y. Given any $r \in R$, it follows from (6) that

$$H[P(y \leq r)] = [P(y \leq r|\, z = 1)P(z = 1),\\ P(y \leq r|\, z = 1)P(z = 1) + P(z = 0)]. \qquad (7)$$

The feasible distribution functions are all increasing functions F(·) such that $F(r) \in H[P(y \leq r)]$ for all $r \in R$.

To go further still, result (7) may be used to obtain sharp bounds on quantiles of y, by inverting the bounds on the distribution function. Manski (1994) and Manski (2003, ch. 1) give alternative derivations of the results for quantiles.

Distributional Assumptions

Distributional assumptions may enable one to shrink identification regions obtained using the empirical evidence alone. One type of assumption asserts that the distribution $P(y|z = 0)$ of missing outcomes lies in some set $\Gamma_{0Y} \subset \Gamma_Y$. Then the identification region shrinks from $H[P(y)]$ to

$$H_1[P(y)] \equiv \left[P(y|\,z=1)P(z=1) + \gamma\left(P(z=0), \gamma \in \Gamma_{0Y}\right]. \qquad (8)$$

Assumptions of this type are not refutable; after all, the empirical evidence reveals nothing about $P(y|z = 0)$. A leading example is the assumption that data are missing at random. Formally, this is the assumption that $P(y|z = 0) = P(y|z = 1)$, which implies that $H_1[P(y)]$ contains the single distribution $P(y|z = 1)$.

A different type of assumption asserts that the distribution of interest, $P(y)$, lies in a set $\Gamma_{0Y} \subset \Gamma_Y$. Then the identification region shrinks from $H[P(y)]$ to

$$H_1[P(y)] = \Gamma_{0Y} \cap H[P(y)]. \qquad (9)$$

Assumptions of the latter type may be refutable: if the intersection of Γ_{0Y} and $H[P\,(y)]$ should be empty, then $P(y)$ cannot lie in Γ_{0Y} For example, let y be real-valued and consider the assumption that $P(y)$ is a symmetric distribution. Then $H_1[P(y)]$ is composed of all members of $H[P(y)]$ that are symmetric. If $H[P(y)]$ contains no symmetric distributions, the empirical evidence reveals that $P(y)$ is not symmetric.

Statistical Inference

The fundamental problem posed by missing data is identification, so it has been convenient in the above discussion to suppose that one knows the distributions that are asymptotically revealed by the sampling process, namely, $P(y|z = 1)$ and $P(z)$. An empirical researcher observing a sample of finite size N must contend with issues of statistical inference as well as identification. I shall not dwell on these here, but merely point out that the empirical distributions $P_N(y|z = 1)$ and $P_N(z)$ almost surely converge to $P(y|z = 1)$ and $P(z)$ respectively. Hence, a consistent estimate of the identification region $H[P(y)]$ is its sample analog

$$H_N[P(y)] = [P_N(y|\,z = 1)P_N(z = 1)\\ +\gamma P_N(z = 0), Y \in \Gamma_Y]. \qquad (10)$$

Moreover, a natural estimate of the identification region for a parameter τ is $\{\tau(\eta),\ \eta \in H_N[P(y)]\}$. Sample analogs may also be used in the presence of distributional assumptions.

Confidence intervals (CIs) may be constructed to measure the sampling variation in estimates of identification regions. Considering cases in which the identification region is an interval on the real line, Horowitz and Manski (2000) propose CIs that asymptotically cover the entire region with fixed probability. Chernozhukov et al. (2004) develop methods for construction of such CIs when the identification region is a general finite-dimensional set. Imbens and Manski (2004) develop a conceptually different confidence interval; rather than cover the entire identification region with fixed probability, their interval asymptotically covers the true value of the parameter with this probability.

Analysis of Treatment Response

Analysis of treatment response poses a pervasive and distinctive problem of missing outcomes. Studies of treatment response aim to predict the outcomes that would occur if different treatment rules were applied to a population. Treatments are mutually exclusive, so one cannot observe the outcomes that a person would experience under all treatments. At most, one can observe the outcome that a person experiences under the treatment he actually receives. The counterfactual outcomes that a person would have experienced under other treatments are logically unobservable.

For example, suppose that patients ill with a specified disease can be treated by drugs or by surgery. The relevant outcome might be lifespan. One may want to predict the lifespans that would occur if all patients were to be treated by drugs. The available data may be observations of the actual lifespans of patients in a study population, some of whom were treated by drugs and the rest by surgery.

To formalize the inferential problem, let each member j of a study population J have a response function $y_j(\cdot)$: $T \rightarrow Y$ mapping the mutually exclusive and exhaustive treatments $t \in T$ into outcomes $y_j(t) \in Y$. Let $z_j \in T$ denote the treatment that person j receives and $y_j = y_j(z_j)$ be the outcome that he experiences. Then $y_j(t)$, $t \neq z_j$ are counterfactual outcomes.

Let $y(\cdot)$: $J \rightarrow Y^{|T|}$ be the random variable mapping the population into their response functions. Let z: $J \rightarrow T$ be the 'status quo treatment rule' mapping the members of J into the treatments that they actually receive. Response functions are not observable, but realized treatments and outcomes may be observable. If so, random sampling from J reveals the status quo (outcome, treatment) distribution $P(y, z)$.

The Selection Problem

Analysis of treatment response seeks to predict the outcomes that would occur under alternatives to the status quo treatment rule. A leading objective is to predict the outcomes that would occur if all persons were to receive the same treatment. By definition, $P[y(t)]$ is the distribution of outcomes that would occur if all persons were to receive a specified treatment t. Hence prediction of outcomes under a rule mandating uniform treatment requires inference on $P[y(t)]$. The problem of identification of this distribution from knowledge of $P(y, Z)$ is commonly called the 'selection problem'.

The selection problem has the same structure as the missing-outcomes problem discussed above. To see this, write

$$\begin{aligned} P[y(t)] &= P[y(t)|\, z=t]P(z=t) \\ &\quad + P[y(t)|\, z \neq t]P(z \neq t) \\ &= P(y|\, z=t)P(z=t) \\ &\quad + P[y(t)|\, z \neq t]P(z \neq t). \end{aligned} \tag{11}$$

The first equality is the Law of Total Probability. The second holds because $y(t)$ is the outcome experienced by persons who receive treatment t. The sampling process reveals $P(y|z = t)$, $P(z = t)$, and $P(z{\neq}t)$, but it is uninformative about $P[y(t)|z{\neq}t]$. Hence, the identification region for $P[y(t)]$ if we use the empirical evidence alone is

$$\begin{aligned} H\{P[y(t)]\} = \{&P(y|\, z=t)P(z=t) \\ &+\gamma P(z \neq t), \gamma \in \Gamma_Y\}. \end{aligned} \tag{12}$$

This identification region has the same form as the region (2) for inference on outcomes with missing data, with $P(z \neq t)$ being the probability of missing data. Hence, all of the analysis of missing outcomes discussed above applies here as well.

Distributional Assumptions

A familiar 'solution' to the selection problem is to assume that the status quo treatment rule makes realized treatments statistically independent of response functions; that is,

$$P[y(\cdot)] = P[y(\cdot)|\, z]. \tag{13}$$

This assumption implies that $P[y(t)] = P(y|z = t)$. The sampling process reveals $P(y|z = t)$. Hence, assumption (13) point-identifies $P[y(t)]$.

Assumption (13) is credible when the status quo treatment rule calls for random assignment of treatments and all persons comply with their assignments. Indeed, the fact that (13) holds is the reason why randomized experiments are held in high esteem. However, the credibility of the assumption in settings without random assignment or full compliance almost invariably is a matter of controversy. This motivates interest in other assumptions that may be better motivated in practice.

There has been much study of assumptions that use an 'instrumental variable'; that is, an observable covariate whose value varies across the study population. Suppose that outcomes are real-valued. Manski (1990) poses the mean-independence assumption $E[y(t)] = E[y(t)|v]$. If outcomes are bounded with values normalized to lie in the unit interval, the resulting identification region for $E[y(t)]$ is

$$H\{E[y(t)]\} = [\max_{v\in V} E\{y \cdot 1[z = t] | v = v\}, \\ \min_{v\in V} E\{y \cdot 1[z = t] + 1[z \neq t] | v = v\}]. \tag{14}$$

Manski and Pepper (2000) study identification of $E[y(t)]$ when v is real-valued and the assumption of mean independence is weakened to state that $E[y(t)|v]$ weakly increases in v. Heckman and Vytlacil (2001) combine the mean-independence assumption with some of the structure of an econometric selection model and show that the identification region for $E[y(t)]$ remains (14).

Statistical independence assumptions are stronger than mean independence. Manski (2003, ch. 7) poses the assumption $P[y(t)] = P[y(t)|v]$ and shows that it yields this identification region for $P[y(t)]$:

$$H\{P[y(t)]\} = \bigcap_{v\in V} \{P(y | v = v,\ z = t) \\ P(z = t | v = v) + \gamma_v \cdot P(z \neq t | v = v),\ \gamma_v \in \Gamma_Y\}. \tag{15}$$

Balke and Pearl (1997) poses the yet stronger assumption $P[y(\cdot)] = P[y(\cdot)|v]$ and characterize its identifying power when outcomes are binary variables.

A different idea, developed in Manski (1995, ch. 6; 1997a) is to place assumptions on the shape of the response functions $y(\cdot)$. One may sometimes believe that treatment response is monotone, in the sense that outcomes increase with the intensity of the treatment. When the set T of treatments is ordered in terms of degree of intensity, the assumption of 'monotone treatment response' asserts that, for all persons j and all treatment pairs (s, t), $t \geq s \Rightarrow y_j(t) \geq y_j(s)$. If outcomes are bounded with values normalized to lie in the unit interval, the resulting identification region for $E[y(t)]$ is the interval

$$H\{Ey(t)\} = [E(y | t \geq z) \cdot P(t \geq z), P(t > z) \\ + E(y | t \leq z) \cdot P(t \leq z)]. \tag{16}$$

A narrower interval results if treatment response is assumed to be concave as well as monotone.

Shape restrictions on the response function and assumptions using instrumental variables illustrate the vast middle ground between inference from the empirical evidence alone and analysis predicated on assumptions that are strong enough to achieve point identification. As the study of partial identification continues to broaden and deepen, empirical researchers will be able to choose from a growing menu of inferential options. One should, however, not expect one uniformly best option to emerge. The appeal of any approach to inference necessarily depends on the objectives of the research, the available data, and the assumptions that are credible to maintain.

See also

- ▶ Non-parametric Structural Models
- ▶ Statistics and Economics
- ▶ Treatment Effect

Bibliography

Balke, A., and J. Pearl. 1997. Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association* 92: 1171–1177.

Blundell, R., A. Gosling, H. Ichimura, and C. Meghir. 2004. Changes in the distribution of male and female wages accounting for employment composition using bounds. In *Working Paper W04/25*. London: Institute of Fiscal Studies.

Bollinger, C. 1996. Bounding mean regressions when a binary regressor is mismeasured. *Journal of Econometrics* 73: 387–399.

Brock, W. 2005. *Profiling problems with partially identified structure. Mimeo*. Madison: WI: Department of Economics, University of Wisconsin-Madison.

Chernozhukov, V., H. Hong, and E. Tamer. 2004. *Parameter set inference in a class of econometric models. Mimeo*. Evanston: IL: Department of Economics, Northwestern University.

Ciliberto, F., and E. Tamer. 2004. *Evanston, IL: Market structure and multiple equilibria in airline markets. Mimeo*. Evanston: IL: Department of Economics, Northwestern University.

Cochran, W. 1977. *Sampling techniques*. 3rd ed. New York: Wiley.

Cochran, W., F. Mosteller, and J. Tukey. 1954. *Statistical problems of the kinsey report on sexual behavior in the human male*. Washington, DC: American Statistical Association.

Cross, P., and C. Manski. 2002. Regressions, short and long. *Econometrica* 70: 357–368.

Dominitz, J., and R. Sherman. 2004. Sharp bounds under contaminated or corrupted sampling with verification, with an application to environmental pollutant data. *Journal of Agricultural, Biological, and Environmental Statistics* 9: 319–338.

Duncan, O., and B. Davis. 1953. An alternative to ecological correlation. *American Sociological Review* 18: 665–666.

Frisch, R. 1934. *Statistical confluence analysis by means of complete regression systems*. Oslo: University Institute for Economics.

Ginther, D. 2002. Alternative estimates of the effect of schooling on earnings. *The Review of Economics and Statistics* 82: 103–116.

Haile, P., and E. Tamer. 2003. Inference with an incomplete model of English auctions. *Journal of Political Economy* 111: 1–51.

Heckman, J., J. Smith, and N. Clements. 1997. Making the most out of programme evaluations and social experiments: accounting for heterogeneity in programme impacts. *Review of Economic Studies* 64: 487–535.

Heckman, J., and E. Vytlacil. 2001. Localinstrumental variables. In *Nonlinear statistical inference: essays in honor of takeshi amemiya*, ed. C. Hsiao, K. Morimune, and J. Powell. Cambridge, MA: Cambridge University Press.

Horowitz, J., and C. Manski. 1995. Identification and robustness with contaminated and corrupted data. *Econometrica* 63: 281–302.

Horowitz, J., and C. Manski. 1998. Censoring of outcomes and regressors due to survey nonresponse: identification and estimation using weights and imputations. *Journal of Econometrics* 84: 37–58.

Horowitz, J., and C. Manski. 2000. Nonparametric analysis of randomized experiments with missing covariate and outcome data. *Journal of the American Statistical Association* 95: 77–84.

Horowitz, J., and C. Manski. 2006. Identification and estimation of statistical functionals using incomplete data. *Journal of Econometrics* 132(2): 445–459.

Horowitz, J., C. Manski, M. Ponomareva, and J. Stoye. 2003. Computation of bounds on population parameters when the data are incomplete. *Reliable Computing* 9: 419–440.

Hotz, J., C. Mullin, and S. Sanders. 1997. Bounding causal effects using data from a contaminated natural experiment: analyzing the effects of teenage childbearing. *Review of Economic Studies* 64: 575–603.

Imbens, G., and C. Manski. 2004. Confidence intervals for partially identified parameters. *Econometrica* 72: 1845–1857.

King, G., and L. Zeng. 2002. Estimating risk and rate levels, ratios and differences in case-control studies. *Statistics in Medicine* 21: 1409–1427.

Klepper, S., and E. Leamer. 1984. Consistent sets of estimates for regressions with errors in all variables. *Econometrica* 52: 163–183.

Koopmans, T. 1949. Identification problems in economic model construction. *Econometrica* 17: 125–144.

Kreider, B., and J. Pepper. 2004. *Disability and employment: reevaluating the evidence in light of reporting errors. Mimeo*. Ames: IA: Department of Economics, Iowa State University.

Manski, C. 1989. Anatomy of the selection problem. *Journal of Human Resources* 24: 343–360.

Manski, C. 1990. Nonparametric bounds on treatment effects. *American Economic Review: Papers and Proceedings* 80: 319–323.

Manski, C. 1994. The selection problem. In *Advances in Econometrics, Sixth World Congress*, ed. C. Sims. Cambridge: Cambridge University Press.

Manski, C. 1995. *Identification Problems in the Social Sciences*. Cambridge, MA: Harvard University Press.

Manski, C. 1997a. Monotone treatment response. *Econometrica* 65: 1311–1334.

Manski, C. 1997b. The mixing problem in programme evaluation. *Review of Economic Studies* 64: 537–553.

Manski, C. 2000. Identification problems and decisions under ambiguity: empirical analysis of treatment response and normative analysis of treatment choice. *Journal of Econometrics* 95: 415–442.

Manski, C. 2001. Nonparametric identification under response-based sampling. In *Nonlinear statistical inference: Essays in honor of Takeshi Amemiya*, ed. C. Hsiao, K. Morimune, and J. Powell. New York: Cambridge University Press.

Manski, C. 2002. Treatment choice under ambiguity induced by inferential problems. *Journal of Statistical Planning and Inference* 105: 67–82.

Manski, C. 2003. *Partial identification of probability distributions*. New York: Springer-Verlag.

Manski, C. 2005a. *Social choice with partial knowledge of treatment response*. Princeton: Princeton University Press.

Manski, C. 2005b. *Search profiling with partial knowledge of treatment response. Mimeo*. Evanston: IL: Department of Economics, Northwestern University.

Manski, C. 2006. Minimax-regret treatment choice with missing outcome data. *Journal of Econometrics* f139(1): 105–115.

Manski, C., and D. Nagin. 1998. Bounding disagreements about treatment effects: a case study of sentencing and recidivism. *Sociological Methodology* 28: 99–137.

Manski, C., and J. Pepper. 2000. Monotone instrumental variables: with an application to the returns to schooling. *Econometrica* 68: 997–1010.

Manski, C., G. Sandefur, S. McLanahan, and D. Powers. 1992. Alternative estimates of the effect of family structure during adolescence on high school graduation. *Journal of the American Statistical Association* 87: 25–37.

Manski, C., and E. Tamer. 2002. Inference on regressions with interval data on a regressor or outcome. *Econometrica* 70: 519–546.

Moinari, F. 2002. *Missing treatments. Mimeo*. Ithaca: NY: Department of Economics, Cornell University.

Molinari, F. 2004. *Partial identification of probability distributions with misclassified data. Mimeo*. Ithaca: NY: Department of Economics, Cornell University.

Pepper, J. 2000. The intergenerational transmission of welfare receipt: a nonparametric bounds analysis. *The Review of Economics and Statistics* 82: 472–488.

Pepper, J. 2003. Using experiments to evaluate performance standards: what do welfare-to-work demonstrations reveal to welfare reformers? *Journal of Human Resources* 38: 860–880.

Peterson, A. 1976. Bounds for a joint distribution function with fixed subdistribution functions: application to competing risks. *Proceedings of the National Academy of Sciences of the United States of America* 73: 11–13.

Reiersol, O. 1941. Confluence analysis by means of lag moments and other methods of confluence analysis. *Econometrica* 9: 1–24.

Robins, J. 1989. The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. In *Health Service Research Methodology: A Focus on AIDS*, ed. L. Sechrest, H. Freeman, and A. Mulley. Washington, DC: NCHSR, US Public Health Service.

Scharfstein, J., C. Manski, and J. Anthony. 2004. On the construction of bounds in prospective studies with missing ordinal outcomes: application to the good behavior game trial. *Biometrics* 60: 154–164.

Stoye, J. 2005. *Partial identification of spread parameters when some data are missing. Mimeo*. Evanston: IL: Department of Economics, Northwestern University.

Zaffalon, M. 2002. Exact credal treatment of missing data. *Journal of Statistical Planning and Inference* 105: 105–122.

Partial Linear Model

Elie Tamer

Abstract

It is popular to summarize the relationship between an outcome variable y and a vector (x, z) through a linear mean regression where the mean of y is modelled as a linear function of both x and z. A more robust specification is called for in some situations where the imposed linear relationship between (the mean of) y and z is suspect. A partially linear specification allows for a regression function that maintains linearity in x but allows the effect of z to be nonlinear. This partially linear model has been widely studied in the statistics and the semiparametric econometrics literature.

Keywords

Age elasticity; Bootstrap; Censored selection models; Conditional expectations; Convergence; Cross validation; Heteroskedasticity; Homoskedasticity; Income elasticity; Kernel estimators; Linear models; Linear regression; Nonparametric estimation; Nonparametric selection models; Partially linear models; Quantile regression; Random variables; Semiparametric estimation; Semiparametric sieve least squares; Sieves; Spline regression

JEL Classifications

C14

A partially linear model requires the regression function to be a linear function of a subset of the variables and a nonparametric non-specified function of the rest of the variables. Suppose, for example, that one is interested in estimating the relationship between an outcome variable of interest y and a vector of variables (x, z). The economist is comfortable modelling the regression function as linear in x, but s hesitant in extending the linearity to z. One example, considered by Engle et al. (1986), is the effect of temperature on fuel consumption using a time series of cities. To do that, one can consider a regression of average fuel consumption in time t on average household characteristic and average temperature in time t. The analyst might be more comfortable with imposing linearity on the part of the regression function involving household characteristics but unwilling to require that fuel consumption varies linearly with temperature. This is natural since fuel consumption tends to be higher at extremes of the temperature scale, but lower at moderate temperatures. The regression function Engle et al. consider is:

$$y = x'\beta + g(z) + u \tag{1}$$

where x denotes a vector of household/city characteristics and z is temperature and u is a mean zero random variable such that is independent of (x, z). The function g(.) is unspecified except for smoothness assumptions. They term this *the semiparametric regression model.*

Another example is the demand for gasoline model used by Schmalensee and Stoker (1999). The primary interest in this paper is the age income structure of household demand for fuel. In particular, the authors want to estimate demand elasticities of age and income (do richer household consume more gasoline, and, if so, by how much?). Hence, their dependent variable is the logarithm of gasoline consumed and $g(.)$ is a function of both age and the logarithm of income. Schmalensee and Stoker also control for a set of other household characteristics. This partially linear model allows one to have a more robust model of the relationship between mean gasoline consumption and age and income.

The partially linear model can arise also as a special case of a censored sample selection model (see Heckman 1974). There, we are interested in estimating β in the equation $y^* = d * (x\beta + \varepsilon)$ where d is a binary observed random variable that indicates censoring: $d = 1$ the outcome y is uncensored, and $d = 0$ otherwise. The model above can be written as

$$E[y|\,w, x, d = 1] = x\beta + g(w).$$

If there is no overlap between x and w, this is an example of a partial linear model with a nonparametric selection mechanism. A more general version of this model is studied in Ahn and Powell (1993).

Partially linear models are more attractive than linear models especially in cases where the linearity assumption on a subset of the regressors is suspect. This more robust model allows for a more flexible parametrization for that part of the regression where the analyst is not convinced of the linearity. On the other hand, the main motivation for modelling the regression function as partially nonparametric, or semiparametric, as opposed to fully nonparametric, is the concern for the precision of the estimates. In particular, with more continuous regressors in the regression the 'curse of dimensionality' slows the rate of convergence, effectively reducing the usefulness of the regression in data-sets with moderate sizes. Hence, partially linear models provide another practical tool for analysts to use in regressions where linearity of part of the regression function is questionable and provides a middle ground between a completely linear regression that is less robust and one that is totally nonparametric but less practical.

There are many approaches to estimating β and g. For example, one can use a penalized spline regression similar to the one used in Engle et al. (1986), or use semiparametric sieve least squares by replacing the function f with an appropriate sieve that approximates the function space (where g lies) as sample size increases. The method we describe here uses kernel smoothing similar to the one used by Robinson (1988) and Speckman (1988). Notice that Eq. 1 above implies that

$$E[y|\,z] = E[x'|\,z]\beta + g(z). \tag{2}$$

Subtracting Eq. 2 from 1 we obtain

$$y - E[y|\,z] = \left(x - E[x|\,z]'\beta + u\right). \tag{3}$$

Hence, one can consistently estimate β by regressing ($y - E[y|z]$) on ($x - E[x|z]$) if the matrix $E[(x - E[x|z])(x - E[x|z])']$ is full rank. This procedure has some similarities to a linear regression where one is interested in a subset of the slope parameters. One can obtain this by regressing the dependent variable on residuals from a regression of the regressors of interest on the nuisance regressors. It is a regression of the outcome on what remains of the regressors after purging them of their linear component that is common with other regressors.

One problem in our set-up is that the regression in Eq. 3 is unfeasible since $E[y|z]$ and $E[x|z]$ are not known. These can be consistently estimated using a variety of methods like kernels or sieves. Robinson (1988), for example, replaces the conditional expectations by appropriate Naradaya–Watson kernel estimators where for a random sample of size N,

$$\widehat{E}[y|\,z] = \frac{1}{N}\sum_{i=1}^{N} w_{in}(z)y_i \tag{4}$$

where the weight function w_{in} is such

$$w_{in}(z) = \frac{K\left(\dfrac{z_i - z}{h_n}\right)}{\dfrac{1}{N}\sum K\left(\dfrac{z_i - z}{h_n}\right)}. \tag{5}$$

K(.) is a kernel function satisfying certain conditions (see Hardle 1991, for more on smoothing conditional expectations), and h_n is a bandwidth parameter that is positive and converges to zero as sample size increases. Conditions on the rate of convergence of this bandwidth are obtained to ensure desirable theoretical properties of the estimators (for example, on the conditional expectation case, we have $h_n = \lambda n^{-1/5}$ where $0 < \lambda < \infty$). Robinson then shows that the estimator $\widehat{\beta}$ of β is normally distributed asymptotically as sample size increases. (The estimator Robinson considers requires trimming those values of z that cause instability in the estimates in the 'random denominator' of the conditional expectation.) In particular,

$$\sqrt{n}\left(\widehat{\beta} - \beta\right) \rightarrow_d \mathcal{N}\left(0,\ \ \sigma^2 E\left[(x - E[x|\,z])\left(x - E[x|\,z]'\right]^{-1}\right). \tag{6}$$

This is derived on the assumption of homoskedasticity ($V(u|x) = a^2$), and other conditions guaranteeing well behaviour of the kernel estimators as sample size increases. As for estimating the nonparametric function g(.), one can use a feasible version of Eq. 2 to get

$$\widehat{g}(z) = \widehat{E}[y|\,z] - \widehat{E}[x|\,z]'\widehat{\beta}. \tag{7}$$

Under a set of assumptions, it can be shown that $\hat{g}\,(z)$ is a consistent estimator for $g\,(z)$. For example, in the case of scalar z that has support on [0,1], it can be shown that under appropriate assumptions,

$$sup_{t\in[0,1]}|\widehat{g}(t) - g(t)|\ \ = O\left(n^{-2/5}\log^{2/5}n\right).$$

In addition, Hardle et al. (2001) provide more consistency results for the nonparametric function g(.).

In practice, to implement a partially linear regression, three additional tasks remain. First, one needs to choose a kernel function. Second, although rates of convergence for the smoothing parameter h_n were given, those provide no guidance for choosing a particular value for this smoothing parameter with a given data-set. Third, and to account for sample variability, one needs to obtain estimates for the variance covariance matrix. As for the choice of the kernel function, one can use $K(u) = (2\pi)^{-1/2}\exp(-\frac{1}{2}u^2)$, $K(u) = \frac{1}{2}1[|\,u| \ \ \le 1]$ or a quartic kernel $15/16((1 - u^2)^2\ I\ (|u| \le 1))$ (see Hardle 1991, for more on kernel selection. Kernel selection does not seem to make a difference in practice). As for the choice of the smoothing parameter h_n, one method that can be used is *cross validation*. In particular, our estimators of β and the function g(.) obtained from Eqs. 3 and 7 can be written as $\widehat{\beta}(h_n)$

and $\hat{g}$ $(z) = \hat{g}$ $(z;\ h_n)$ which are functions of the smoothing parameter h_n. So, to choose h_n in practice, one can minimize the cross-validation function $cv(h_n)$ defined as

$$cv(h_v) = \frac{1}{N}\sum_{i=1}\left({}_i - x_i\widehat{\beta}(h_n) - \widehat{g}(z_i; h_n)\right)^2. \quad (8)$$

Finally, to estimate the variance covariance matrix in the homoskedastic case, one can replace $\sigma^2 E[(x - E[x|\ z])\ (x - E[x|\ z]')]^{-1}$ by its sample analog. In particular, an estimator $\widehat{\sigma^2}$ of σ^2 can be:

$$\widehat{\sigma^2} = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - x_i\widehat{\beta} - \widehat{g}(z_i)\right)^2.$$

However, since the conditional mean is semiparametric, a better estimator for the variance matrix is one that is heteroskedasticity robust. This estimator is similar to the heteroskedasticity robust estimator in linear regression and can be written as

$$\widehat{V} = \frac{1}{N}\sum\left(x_i - \widehat{E}(x|\ z_i)\right)\ \left(x_i - \widehat{E}(x|\ z_i)\right)' \left(y_i - \widehat{\beta}x_i - \widehat{g}(z_i)\right)^2.$$

One can also approximate the finite sample distribution of the estimator by a bootstrapped distribution. After estimating β and $g(.)$, one obtains a set of centred residuals $e_i^*, i = 1, \ldots, N$ with distribution F_n^* from which one can draw a bootstrap sample, and then generate a sample of y's from which one can obtain one's bootstrap estimates. Hardle et al. (2001) contains consistency results for the bootstrap procedure in the partially linear model above.

Partially linear models are semiparametric linear regressions where the regression function contains a nonparametric function. These regressions are robust to the linear specification for part of the regressors. In addition, partially linear models provide a good alternative to fully nonparametric regression in settings where the data-set that is available is of moderate sample sizes and/or when one has to smooth over a set of continuous random variables of high dimension. Finally, one can also extend the independence (or mean independence) usually used in estimating partially linear models to conditional quantile restrictions and obtain a partially linear semiparametric quantile regression.

See Also

- ▶ Semiparametric Estimation

Bibliography

Ahn, H., and J. Powell. 1993. Semiparametric estimation of censored selection models. *Journal of Econometrics* 58: 3–29.

Engle, R., C. Granger, C. Rice, and J. Weiss. 1986. Semiparametric estimates of the relation between weather and electricity sales. *Journal of the American Statistical Association* 81: 310–320.

Hardle, W. 1991. *Applied nonparametric regression*. New York: Cambridge University Press.

Hardle, W., H. Liang, and J. Gao. 2001. *Partially linear models (contributions to statistics)*. Heidelberg: Physica-Verlag.

Heckman, J. 1974. Shadow wages, market wages and labor supply. *Econometrica* 42: 679–693.

Robinson, P. 1988. Root-n-consistent semiparametric regression. *Econometrica* 56: 931–954.

Schmalensee, R., and T. Stoker. 1999. Household gasoline demand in the United States. *Econometrica* 67: 645–662.

Speckman, P. 1988. Kernel Smoothing in partial linear models. *Journal of the Royal Statistical Society, B* 50: 413–436.

Pascal, Blaise (1623–1662)

A. W. F. Edwards

Keywords

Expectations; Fermat, P. de; Pascal, B.; Probability

JEL Classifications

B31

Pascal was born on 19 June 1623 in Clermont, France, and died on 19 August 1662. In 1631 his father Etienne Pascal moved to Paris in order to secure his son a better education. In 1635 Etienne was one of the founders of Marin Mersenne's 'Academy', to which he introduced his son at the age of 14, and Blaise immediately put this new source of knowledge to good use, producing (at the age of 16) his famous *Essai pour les coniques.*

In the succeeding years the young Pascal designed and had built the first mechanical adding machine (there is now a computer language called 'Pascal') and conducted experiments into the nature of a vacuum (the 'Pascal' is the S.I. unit of pressure), but his chief mathematical contribution was to lay the foundations of the theory of probability.

Before his time probability calculations amounted to no more than the enumeration of equally probable outcomes in games of chance, but Pascal introduced the important idea of expectation and used recursively the fact that if expectations of gain X and Y are equally probable, the expectation is $\frac{1}{2}(X+Y)$. He also introduced the binomial distribution for equal chances and with its help, and that of mathematical induction applied to expectations, solved the Problem of Points for two players.

This problem was the topic of correspondence between Pascal and Pierre de Fermat in 1654 which, together with Pascal's contemporary *Traité du triangle arithmétique,* includes three methods of solution. Two players stake equal money on being the first to win n points in a game in which the winner of each point is determined by the toss of a coin. If such a game is interrupted when one player still lacks a points and the other b, how should the stakes be divided between them?

Fermat and Pascal independently concluded that the problem could be solved by noting that at most $(a + b - 1)$ more tosses will settle the game, and that if this number of tosses is imagined to have been made, the resulting 2^{a+b-1} possible games (each equally probable) may be classified according to the winner in each case, the stakes then being divided accordingly. Thus the real game, of indeterminate length, is embedded in an imaginary game of fixed length. Apart from this novel idea, however, such a solution by enumeration was straightforward, but Pascal offered both an independent method based on expectations which is valid for any number of players, and, in the *Traité du triangle arithmétique,* the solution for two players in terms of the binomial distribution, proved by induction. He did not give the binomial distribution algebraically, but by reference to the 'arithmetical triangle' of binomial coefficients, whose properties he elaborated in his *Traité* (whence the name 'Pascal's triangle').

In the *Pensées* Pascal introduced his celebrated wager '*infini-rien*', in which he argued that we should wager for the existence of God since the stakes are finite (our lives) but there is an infinite prize (eternal life). The argument is that of modern decision theory, which it may be said to foreshadow. The 'states of nature' are the existence and non-existence of God whilst the 'decisions' are to act as if God exists and as if he does not. If God does exist and we act as if he does, the 'utility' is infinite, and thus so is the 'expected utility' of this course of action whether he exists or not, provided there is a non-zero chance that he does.

Patent Pools

Daniel Quint

Abstract

A patent pool is an agreement by multiple patentholders to share intellectual property among themselves or to license a portfolio of patents as a package to outsiders. Patent pools were common in the United States from the 1890s to the 1940s; since the mid-1990s, there has been a resurgence of patent pools tied to technological standards. I discuss the history and antitrust treatment of patent pools in the United States, and review the related academic literature (both theoretical and empirical).

Keywords

Antitrust legislation; Complements; Intellectual property; Patent pools; Patents

JEL Classifications

K21; L24; L4; O34

A patent pool is an agreement by multiple patentholders to share intellectual property among themselves or to license a portfolio of patents as a package to outsiders.

Patent pools were common in the United States in the first half of the 20th century, and reemerged as an important institution in the mid-1990s; an estimated $100 billion worth of goods sold in 2001 were based at least partly on pooled patents.

History

The first patent pool emerged from infringement lawsuits won by Elias Howe, credited with inventing the sewing machine, who returned from marketing his invention in England in the 1840s to find that others had copied it. Following the lawsuits, Howe, Isaac Singer and two other manufacturers established a pool of sewing machine-related patents in 1856, with Howe receiving the bulk of the royalties.

Patent pools were commonplace in the United States from the 1890s to the 1940s. Lerner et al. (2007) identify 125 pools, most of them from this time; Lerner and Tirole (2007) claim that in the early 20th century, 'many (if not most) important manufacturing industries had a patent pooling arrangement'. (A partial list from Merges (2001) includes pools covering shoe machinery, automobiles, bathtubs, door parts, seeded raisins, coaster brakes, davenport beds, movie projectors, hydraulic pumps, and swimming pool cleaners; a longer list from Lerner et al. includes railroad couplers, television equipment, and plastic artificial eyes.) In 1917, with airplanes needed for the First World War, then Assistant Secretary of the Navy Franklin D. Roosevelt pushed eight aircraft manufacturers into a patent pool because patent litigation had shut down US aircraft production. A 1915 pool containing automobile patents had 146 initial members, but most of the pools examined in Lerner, Strojwas and Tirole started with six members or fewer.

Following Congressional hearings on patent pools in the 1930s and 1940s and several negative antitrust rulings, patent pools essentially vanished from the mid-1950s until the mid-1990s. In 1997, after extensive discussion with regulators, a pool formed containing patents essential to the MPEG-2 digital video standard. This was followed by pools tied to the DVD, Bluetooth, 1394 (Firewire), DVB-T, MPEG-4 (AVC) and 3G–Mobile standards. The MPEG-2 pool alone currently has 26 members, nearly a thousand patents, and over 1300 licensees and affiliates. Pools have also recently been discussed for the biotech and pharmaceutical industries.

Antitrust Treatment

For two decades following the passage of the Sherman Antitrust Act in 1890, patent pools appeared to offer a way to circumvent its prohibitions. In 1902, the Supreme Court upheld the legality of the National Harrow pool, which dominated the market for float spring tooth harrows. Among other things, the licensing terms required licensees to only sell particular products, and fixed the prices for these products. The Court wrote:

> The general rule is absolute freedom in the use or sale of rights under the patent laws of the United States. The very object of these laws is monopoly, and the rule is, with few exceptions, that any conditions which are not in their very nature illegal with regard to this kind of property… will be upheld by the courts. (E. Bement & Sons *v.* National Harrow *(186 US 70)*)

In 1912, however, the Court reversed itself, upholding a lower court's break-up of a pool with similarly restrictive licensing terms (*Standard Sanitary Manufacturing* v. *United States*). In the decades following, the court continued to focus on licensing terms, breaking up pools that fixed downstream prices or production, and allowing pools whose licensing agreements 'contained no restrictions as the quantity of goods to be produced, or the price to be charged, or the territory in which they might be sold by the licensee' *(Baker-Cammack Hosiery Mills* v. *Davis,* 181 F.2d 550 1950). In 1945, the Supreme Court ruled against the Hartford-Empire pool, which used licensing terms to set production quotas in the glassware manufacturing industry, claiming, 'The history of this country has perhaps never witnessed a more completely successful economic tyranny over any field of industry' *(Hartford Empire Co.* v. *United States*, 323 US 386). Although the Baker-Cammack ruling followed that, several other pools were broken up in subsequent years (*United States* v. *Line Material, United States* v. *U.S. Gypsum, United States* v. *New Wrinkle*), and Hartford-Empire was generally seen as signalling the end of favourable treatment toward pools; by the mid-1950s, pool formation had essentially ceased.

This changed following release of the Antitrust Guidelines for the Licensing of Intellectual Property by the Department of Justice and Federal Trade Commission in April 1995. Under the heading 'cross-licensing and pooling arrangements,' the Guidelines stated:

> These arrangements may provide procompetitive benefits by integrating complementary technologies, reducing transaction costs, clearing blocking positions, and avoiding costly infringement litigation. By promoting the dissemination of technology, cross-licensing and pooling arrangements are often procompetitive.

Department of Justice analysis, enunciated in business review letters of several proposed pools, focused on three questions: whether a pool would integrate complementary patent rights (as opposed to patents which would otherwise be in competition); whether it would foreclose competition in related markets; and whether it would discourage further innovation. In the cases of the MPEG-2, DVD, and 3G pools, the DOJ stated after review that it was 'not presently inclined to initiate antitrust enforcement action against the conduct you have described'. In 1998, the FTC did challenge a pool formed by Summit Technology and VISX, the only firms with FDA-approved technology for laser eye surgery, which was viewed to be functioning primarily as a price-fixing arrangement; the pool was dissolved as part of a settlement resolving the case. A 2007 DOJ/FTC report, which followed public hearings held in 2002, summarizes the current regulatory view.

Characteristics of Recent Pools

To address the first regulatory concern – the integration of only complementary patent rights – recent pools have been limited to patents deemed essential for standard compliance. The business review letter on the proposed MPEG-2 pool reads:

> The Portfolio combines patents that an independent expert has determined to be essential to compliance with the MPEG-2 standard; there is no technical alternative to any of the Portfolio patents within the standard. Moreover, each Portfolio patent is useful for MPEG-2 products only in conjunction with the others. The limitation of the Portfolio to technically essential patents, as opposed to merely advantageous ones, helps ensure that the Portfolio patents are not competitive with each other. . .. The continuing role of an independent expert to assess essentiality is an especially effective guarantor that the Portfolio patents are complements, not substitutes. (Joel Klein (Acting Assistant Attorney General), letter to Garrard Beeney, 26 June 1997.

Several of the recent pools include grantback provisions – pool participants and licensees agree to add to the pool, or to license to each other at reasonable terms, any future patents they receive that are judged to be essential. The pools also allow for separate licensing of individual patents – that is, licensing through the pool is not done exclusively. The majority of the recent pools allocate revenue in proportion to the number of essential patents that each firm has contributed to the portfolio, although some of the pools do attempt to account for patents that are more or less valuable.

One unusual case is the 3-G mobile standard. 3-G was designed to use five different radio interfaces, in order to be backward-compatible with five second- generation wireless networks. Antitrust concerns led to the establishment of five separate License Administrators to oversee licensing of patents essential for each interface, rather than a single platform or pool containing all of the relevant patents. (The 3-G platforms are different from traditional pools in that all licensing is done ‘a la carte’, at standardized terms set by each Administrator.)

Theoretical Literature

Shapiro (2001) employs a Nash-Bertrand model to show that pools result in lower prices and greater welfare when patents are perfect complements, by correcting the ‘complements problem’ of excessive prices; and higher prices and lower welfare when patents are perfect substitutes, by eliminating competition. Kim (2004) finds that when patents are perfect complements, the case for pools is even stronger in the presence of vertically integrated firms (patentholders who are also downstream producers). Choi (2003), on the other hand, shows that patent pools change the incentive for another patentholder or a potential infringer to challenge questionable patents in court, making pools of complementary but weak patents possibly welfare- destroying.

Lerner and Tirole (2004) introduce a more flexible model than perfect complements and perfect substitutes, and show that when patents are more substitutable, pools are more prone to be welfare-negative. They show that forcing pool participants to also make their patents available individually has a destabilizing effect on welfare-negative pools, but no effect on welfare-positive pools, and therefore propose compulsory individual licensing as a screen for efficient pools. Brenner (forthcoming) examines the equilibrium effects of different pool formation rules in the Lerner and Tirole framework, showing that endogenously occurring pools will be inefficiently small if patentholders can opt out individually without disrupting pool formation. My own work (Quint 2008) examines pools in a setting with both essential and nonessential patents; I find that pools of essential patents are always welfare-increasing, while pools containing nonessential patents have ambiguous welfare effects, even when they are limited to patents that are perfect complements. I also find that when a pool is welfare-increasing, agreements that ‘bind the pool’s hands’ with respect to pricing will reduce, and may even reverse, the welfare gains.

Empirical Literature

Merges (2001) discusses the workings of many historical pools. Gilbert (2004) discusses a number of important court rulings and how they hold up under economic analysis. Lerner et al. (2007) analyse the licensing rules of 63 patent pools, most from before 1950 but a handful from the 1990s; they find that, consistent with theory, pools containing complementary patents were more likely to allow independent licensing and require grantbacks. Layne-Farrar and Lerner (2008) examine arrangements for dividing pool revenue and its effect on participation; they also find that vertically integrated firms are more likely to join pools. Lerner and Tirole (2007) review current public policy and suggest certain changes.

See Also

- ▶ Anti-trust Enforcement
- ▶ Intellectual Property
- ▶ Patents

Bibliography

Brenner, S. Forthcoming. Optimal formation rules for patent pools. *Economic Theory.*

Choi, J.P. 2003. Patent pools and cross-licensing in the shadow of patent litigation. CESifo working paper #1070. Available at: http://ssrn.com/abstract=466062

Gilbert, R. 2004. Antitrust for patent pools: A century of policy evolution. *Stanford Technology Law Review, 3.*

Kim, S. 2004. Vertical structure and patent pools. *Review of Industrial Organization* 25(3): 231–250.

Layne-Farrar, A., and J. Lerner. 2008. To join or not to join: Examining patent pool participation and rent sharing rules. SSRN working paper. Available at: http://ssrn.com/abstract=945189

Lerner, J., and J. Tirole. 2004. Efficient patent pools. *American Economic Review* 94(3): 691–711.

Lerner, J., and J. Tirole. 2007. Public policy toward patent pools. In *Innovation policy and the economy*, ed. A. Jaffe, J. Lerner, and S. Stern, vol. 8, 157–186. Chicago: University of Chicago Press.

Lerner, J., M. Strojwas, and J. Tirole. 2007. The design of patent pools: The determinants of licensing rules. *RAND Journal of Economics* 38(3): 610–625.

Merges, R. 2001. Institutions for intellectual property transactions: The case of patent pools. In *Expanding the boundaries of intellectual property: Innovation* *policy for the knowledge society*, ed. R. Dreyfuss, D. Zimmerman, and H. First, 123–166. Oxford: Oxford University Press.

Quint, D. 2008. Economics of patent pools when some (but not all) patents are essential. Working paper. Available at: http://www.ssc.wisc.edu/~dquint/papers/patent-pools-quint.pdf

Shapiro, C. 2001. Navigating the patent thicket: Cross licenses, patent pools, and standard-setting. In *Innovation policy and the economy*, ed. A. Jaffe, J. Lerner, and S. Stern, vol. 1, 119–150. Cambridge, MA: MIT Press.

Regulatory Guidelines and Business Review Letters

Letter from Charles James, Assistant Attorney General, Department of Justice, to Ky Ewing, 12 November 2002. Available at: http://www.usdoj.gov/atr/public/busreview/200455.pdf

Letter from Joel Klein, Acting Assistant Attorney General, Department of Justice, to Garrard Beeney, 26 June 1997. Available at: https://www.justice.gov/sites/default/files/atr/legacy/2006/10/17/215742.pdf

Letter from Joel Klein, Assistant Attorney General, Department of Justice, to Garrard Beeney, 16 December 1998. Available at: http://www.usdoj.gov/atr/public/busreview/ 2121.htm

Letter from Joel Klein, Assistant Attorney General, Department of Justice, to Carey Ramos, 10 June 1999. Available at: http://www.usdoj.gov/atr/public/busreview/2485.htm

US Department of Justice and Federal Trade Commission. 1995. *Antitrust guidelines for the licensing of intellectual property*. Available at: http://www.usdoj.gov/atr/public/guidelines/0558.htm

US Department of Justice and Federal Trade Commission. 2007. *Antitrust enforcement and intellectual property rights: Promoting innovation and competition*. Chapter 3: Antitrust analysis of portfolio cross-licensing agreements and patent pools. Available at: http://www.usdoj.gov/atr/public/hearings/ip/chapter_3.htm

Patent Races

Richard A. Jensen

Abstract

A patent race is a competition between two or more inventors, typically firms, to discover an invention first, thereby securing a patent which protects the invention from imitation. The date at which a firm discovers the invention is stochastic, but can be reduced in expectation by increased investment in research and development. Competition to win the patent leads firms to over-invest, compared with the outcome where they invest cooperatively and share the patent equally. However, the expected discovery date is later than socially optimal, so innovation is delayed, on average, compared with the social optimum.

Keywords

Patents; Research and development; Exponential distribution; Schumpeterian hypothesis; Hazard function; Nash equilibrium; Feedback strategies; Subgame perfect equilibrium

JEL Classifications

O3

A patent race is a situation in which two or more inventors, typically firms, compete to discover an invention first, thereby securing a patent which protects the invention from imitation or infringement.

The literature on patent races is predominantly theoretical. These analyses have two fundamental properties. First, for each firm, the discovery date of the invention is stochastic, and depends on the effort expended (or investment) by both itself and its rivals. It is common to assume the discovery date is a random variable that is exponentially distributed with a parameter that depends on the knowledge levels of the firms, which in turn depend on the cumulative research and development (R&D) investments of the firms. If firm i invests an amount $I_i(t)$ at date t, then the growth of its knowledge stock $K_i(t)$ is $K_i'(t) = I_i(t)$. The distribution of firm i's random discovery date t_i is $Pr\{t_i \leq t\} = F(K_i(t)) = 1 - exp\{-\lambda K_i(t)\}$, where $\lambda > 0$ is a parameter, noteworthy because the cumulative knowledge needed for discovery is exponentially distributed with mean $1/\lambda > 0$. The probability that i discovers at t, conditional on i not discovering before t, is $Pr\left\{t_i \in \left(t, t + dt\right] | t_i > t\right\} = \lambda I_i(t) dt$.Given n firms, if their research processes are stochastically independent, then the probability density that i wins the race at t is $\lambda I_i(t) exp\left\{-\lambda \sum_{j=1}^{n} K_i(t)\right\}$.

Second, the race is modelled as a game in which each firm chooses its R&D investment (effort) at each t to maximize the present discounted value of its expected profit, subject to knowledge growth $K_i'(t) = I_i(t)$ from an initial stock $K_i(0) = K_0 \geq 0$. If $V > 0$ is the value of the invention at its discovery date, $r > 0$ is the interest rate, and $c_i(x_i)$ is i's R&D cost function, then this expected present value is

$$\int_0^\infty exp\{-rt\} \times \left[\lambda I_i(t) exp\left\{-\lambda \sum_{j=1}^{n} K_i(t)\right\} V - exp\left\{-\lambda \sum_{j=1}^{n} K_i(t)\right\} c_i(x_i(t))\right] \times dt.$$

Research on patent races initially focused on the Schumpeterian hypothesis regarding the relationship between competition and the pace of innovative activity. In the seminal article on patent races, Loury (1979) assumes each firm chooses a lump-sum R&D expenditure at the start, so $c_i(x_i(t)) = x_i(t)$ where $x_i(0) = X_i$ and $x_i(t) = 0$ for $t > 0$. With no knowledge accumulation over time, the probability density of discovery becomes $h(X_i)$ $exp\left\{-\sum_{j=1}^{n} h(X_i)t\right\}$ where the hazard function $h(X_i)$, the probability that i discovers at t, given that it has not discovered before t, depends only on the lump-sum R&D expenditure. Thinking of invention as a stochastic production process, it is natural to assume that this hazard function is increasing in expenditure, possibly with initially increasing returns to scale, but necessarily with decreasing returns eventually. In this model, increased competition (an increase in the number of firms) reduces the Nash equilibrium expenditure of each firm. Given a fixed number of firms, however, each firm spends more than in the outcome where they invest cooperatively and share the patent value equally. Thus, with unrestricted entry, there are too many firms and too much aggregate R&D investment in the Nash equilibrium, compared with the cooperative outcome.

Lee and Wilde (1980) note that Loury's approach does not allow firms to invest in R&D over time. They assume instead that each firm initially chooses a level of R&D expenditure for each date until it or a rival discovers the invention, $c_i(x_i(t)) = x_i$ for all $t \geq 0$ before discovery and $c_i(x_i(t)) = 0$ thereafter. Again, with no knowledge accumulation, the probability density of discovery is $h(x_i)exp\left\{-\sum_{j=1}^{n} h(x_i)t\right\}$ where the hazard function $h(x_i)$ now depends only on current R&D expenditure. In this case, increased competition increases the Nash equilibrium expenditure of each firm.

For a fixed number of firms, each firm spends more than in the outcome where they invest cooperatively and share the patent value equally. And with unrestricted entry, there are too many firms, each of which spends too much on R&D, in the Nash equilibrium, compared with the cooperative outcome.

One notable difference between these studies is that competition increases R&D investment per firm in Lee and Wilde's approach. This arises from the different R&D strategies. In Loury's model, firms choose the scale of R&D effort with one initial investment, whereas in Lee and Wilde's model, firms choose the intensity of R&D effort per period. In the latter approach, firms can cut their R&D spending, and so their losses, after a rival discovers.

Both of these patent races are essentially static in that the firms choose the strategies at the beginning of the game, and there is no knowledge accumulation. Reinganum (1982) generalizes this by allowing firms to choose feedback strategies: each firm chooses its R&D investment at each date as a function of the observed knowledge stocks of all firms in the race. When patent protection is perfect, increased competition increases the R&D investment expenditure of each firm. However, the effect of increased competition on a firm's R&D investment is ambiguous when patent protection is not perfect. Finally, when the social value of the patent exceeds the private value to a firm in the race, the noncooperative growth rate of knowledge is less than socially optimal, and so innovation is delayed on average compared with the social optimum.

Subsequent research has sought to understand the effects of two types of asymmetries on patent race outcomes. The first type involves an incumbent who owns the current patent and a group of potential entrants vying for the patent to the next generation technology. There are two conflicting effects. First, there is a dissipation of monopoly rent if the incumbent loses. If the innovation is a non-drastic new process, and $M =$ monoply profit and $D =$ duopoly profit, then the incumbent earns $M - D$ if it wins, and the entrant earns D, where typically $M - D > D$. However, if the monopolist wins, it replaces itself as the monopolist (Arrow 1962). If the innovation is drastic and pre-innovation profit is π, then the monopolist's gain from winning is $M - \pi$, while the entrant's is $M > M - \pi$. Incumbents have a greater incentive to innovate when the rent dissipation effect dominates (Gilbert and Newbery 1982), but not when the replacement effect dominates (Reinganum 1983).

The second asymmetry involves a race in which one firm has a lead in the race. When the lead takes the form of greater accumulated knowledge, $K_i(0) > K_j(0)$ for all $j \neq i$, the laggards simply exit and concede the race to the leader firm i in the unique subgame perfect equilibrium. However, the laggards can and do remain in the race if there is some way that they can leapfrog into the lead, such as when the R&D process requires completion of several successive stages (Fudenberg et al. 1983).

See Also

- ▶ Intellectual Property
- ▶ Patents

Bibliography

Arrow, K. 1962. Economic welfare and the allocation of resources for invention. In *The rate and direction of inventive activity*, ed. R.R. Nelson. Princeton: Princeton University Press.

Fudenberg, D., R. Gilbert, J. Stiglitz, and J. Tirole. 1983. Preemption, leapfrogging and competition in patent races. *European Economic Review* 22: 3–31.

Gilbert, R., and D. Newbery. 1982. Preemptive patenting and the persistence of monopoly. *American Economic Review* 72: 514–526.

Lee, T., and L.L. Wilde. 1980. Market structure and innovation: A reformulation. *Quarterly Journal of Economics* 94: 429–436.

Loury, G.C. 1979. Market structure and innovation. *Quarterly Journal of Economics* 93: 395–410.

Reinganum, J.F. 1982. Dynamic game of R and D: Patent protection and competitive behavior. *Econometrica* 50: 671–688.

Reinganum, J.F. 1983. Uncertain innovation and the persistence of monopoly. *American Economic Review* 73: 741–748.

Patent Valuation

Jonathan Putnam

Abstract

Agents routinely appraise and trade individual patents. But small-sample methods (generally derived from basic accounting and finance) are often crude, and their results may bear little relationship to economic fundamentals, especially in litigation. Meanwhile, large-sample methods usually lack much invention-specific data on which to condition value estimates. Regardless of sample size, proper valuation methods require both conceptual delineation and empirical ingenuity.

Keywords

Disclosure; Inventions; Patent citations; Patent counts; Patent valuation; Patents; Research and development

JEL Classification

O31

The valuation of patent rights sounds like a simple enough concept. It is true that agents routinely appraise and trade individual patents. But small-sample methods (generally derived from basic accounting and finance) are often crude, and their results may bear little relationship to economic fundamentals, especially in litigation. On the other hand, large-sample methods usually lack much invention-specific data on which to condition value estimates. Regardless of sample size, proper valuation methods require both conceptual delineation and empirical ingenuity.

Concepts

Legally, a patent is the right to exclude others from making, using or selling an invention. In economic terms, that right is an asset, yielding a non-negative returns stream while it is enforceable. Because the right is a private means (increased exclusivity) to a public end (increased productivity), a patent's private value only partially conveys its market significance.

Unlike most property rights, patents do not comprise the affirmative right to use the invention. Absent the right to use, patents may generate private value only when combined with complementary assets, such as a licence under other patents. Contracting problems (for example, asymmetric information) may strongly influence value.

A patent may generate private returns apart from the right to exclude rivals. The patentee may use it: to monitor employee performance; to signal otherwise unobservable quality to prospective financiers; to enhance reputation; to signal a willingness to litigate; or to reduce the costs of settlement in the event that litigation occurs ('defensive patenting'). In large samples, it is usually impossible either to observe the magnitude and timing of these sources of value, or to decompose them.

Patents also impose unobservable private costs on the patentee. Chiefly, the inventor must disclose the means for reproducing the invention. Disclosure reduces the cost to rivals of reproducing the invention (static spillover) and conducting R&D (dynamic spillover). Apart from reducing the incentive to invent, these private costs imply social benefits not captured by the patentee.

Cross-sectionally, patents are usually modelled as having a one-dimensional 'quality' (which is either synonymous with, or a monotone function of, the patent's value). More precisely, a patent's private value depends significantly on the exclusivity conferred by its claims, but its uncaptured social value depends significantly on the scope of its disclosure (which must be at least as broad as the claims). For various reasons, including rival use of the patentee's disclosure to develop competing innovations ('creative destruction'), the social and private values of a patent may diverge. Thus, it is theoretically preferable, but empirically much less tractable, to model patents as having two-dimensional 'quality'.

Over time, because of ongoing research by the patentee and his rivals, the private returns to patent protection may fluctuate sharply up or down, in response to complementary or competitive discoveries. The variance is likely to be larger in a patent's early years.

Stylized Facts

The following stylized facts bear on the calculation of aggregate private patent values:

1. Whether aggregated by firm, industry or country, patent counts do not vary much from one period to the next.
2. The distribution of patent values is skewed.
3. Social and private patent values are imperfectly correlated.
4. *Ex ante* and *ex post* values are imperfectly correlated.
5. Most patents are not traded.
6. Samples are selected (not all innovations are patented; not all applications are filed in any single country; not all applications are granted).

Related Research

Proceeding in the direction of generally increasing complexity and structure, the following categories describe large-sample models that economists have developed to value patent rights. Lanjouw et al. (1998) surveys recent papers.

Patent Counts

A variety of models employ simple patent counts to indicate the value of patent rights. Strictly speaking, patent counts indicate quantities rather than values. Under certain assumptions, relative quantities may be proportional to relative values. For example, if two patent samples are drawn from the same value distribution, then the ratio of quantities is an efficient estimator of the ratio of values.

Griliches (1990) reviews a large number of studies that, implicitly or explicitly, rely on this assumption. Griliches' view of 'patent [counts] as economic indicators' is not encouraging ('The food here is terrible.' 'Yes, and the portions are so small.'). Stylized facts 1 and 2 combine to thwart inference. A firm facing a fixed budget constraint may patent its best *N* inventions, which implies little intertemporal variation in patent counts even if their realized quality varies markedly. Thus, patent counts are a *biased* measure of value. Because R&D outcomes are highly variable and skewed, patent counts are an *imprecise* measure of value. For these reasons, the assumption that patent samples are drawn from the same distribution is difficult to test, and often false.

On the other hand, fixed budget constraints for R&D and patenting imply that patent counts may proxy for the value of R&D *inputs*. Hausman et al. (1986) model the lag relationship between patent counts and R&D, and find an approximately contemporaneous relationship.

One may compute implied patent values by associating patent counts with other observable aggregates. On the macro level, McCalman (2005) employs the structural imitation model of Eaton and Kortum (1996) to determine international 'trade' in patents. He estimates that the worldwide value of patent applications filed by US inventors in 1988 was about $12.4 billion ($163,700 per application). The estimates for four other large patenting countries vary: France, $147,200; Germany, $82,200; UK, $53,100; Japan, $47,700.

At the firm level, Pakes (1986) constructs a time series model of patent applications, R&D and the stock market rate of return. Controlling for R&D expenditures, an unanticipated patent application implies an $800,000 increase in market capitalization. This relatively high value also reflects investors' revised expectations of research success, and the selection of publicly traded patentees (which are larger and more successful than average).

Patent Citations (Weighted Patent Counts)

Patent examiners cite prior patents when they decide whether to grant a patent application. Analysts count these citations to indicate the value of the cited patent. Patent counts are then weighted by the number of citations. A recent book-length treatment is Jaffe and Trajtenberg (2002).

This branch of the literature divides in two: estimates of the relationship between citations and patent value; and studies that assume that relationship. In the former category, Trajtenberg's (1990) pioneering study showed that citation-weighted patent counts perform better than unweighted counts in explaining aggregate patent value (see Harhoff et al. 1999). However, this and subsequent studies found that citations tend to indicate the social value of the patent rather than the purely private value (stylized fact 3). Private value is better captured by 'self-citations' from the patentee's own later inventions. Hall et al. (2005) show that weighted patent counts are associated with – and predict – higher stock market returns.

Assuming that citations proxy for value, Henderson et al. (1998) examine the contribution of university patenting to commercial technology; Trajtenberg et al. (1997) find that the 'basicness' of university patents relative to corporate patents has narrowed over time. Jaffe et al. (1993) model the spatial distribution of dynamic spillovers.

Other Indicator-Based Methods

Lanjouw and Schankerman (2004) construct a composite index of patent quality using several indicators (forward and backward citations, number of claims, and number of filing countries). This combination of *ex ante* and *ex post* measures (stylized fact 4) efficiently aggregates informationally distinct components of patent value. The composite also explains related *ex post* decisions (for example, patent renewal and litigation); forward citations (an *ex post* measure) demonstrate the greatest explanatory power.

Structural Models: Patent Renewals and Patent Applications

Although most patents are not traded (stylized fact No. 5), patent office rules effectively require patentees to make optimal investments to create and maintain patent rights. These investments reveal information about the expected value of the asset. The information is censored, however, because (conditional on choosing to invest) patentees make the same investment regardless of the expected value. Structural econometric models identify the underlying value distribution.

Most countries require that a patentee pay an increasing fee to keep a patent right in force. Beginning with Pakes and Schankerman (1984), so-called patent renewal models exploit the optimal stopping problem implicit in the annual investment decision. The *ex post* value distribution is identified from the shares of an annual cohort that are renewed each subsequent year when patentees confront known renewal fee schedules, observed over multiple cohorts. In relatively simple *deterministic* models (Schankerman and Pakes 1986; Sullivan 1994; Schankerman 1998), returns are assumed to depreciate at a known rate following an initial draw from the value distribution. In more complex *options* models (Pakes 1986; Lanjouw 1998), returns evolve stochastically. In both models, the average patent value is relatively low (for example, less than $20,000 in Europe during the post-war period). Lorenz plots reveal that the top 10% of patents account for about 47% of the total value distribution.

The value distribution may also be identified from cross-sectional information (Putnam 1996). Under international rules, patent applicants typically determine simultaneously whether to file in each jurisdiction outside their home jurisdiction. Applicants file if the capitalized value of net returns exceeds the application cost. Application models capture filing anywhere in the world, conditional on a common information set, which mitigates both intertemporal (stylized fact #4) and sample selection (stylized fact #6) problems. The *ex ante* value distribution is identified from the combination of filing countries, assuming that national returns are the product of a common invention-level 'random effect' and an idiosyncratic national market draw. Putnam (1996) values the mean German patent at about $69,000 in 1974, with the top 10% of patents accounting for about 70% of the value distribution.

Small-Sample Methods

Small-sample patent valuation typically occurs in a legal or quasi-legal context, such as licensing or litigation. In infringement litigation, the law typically allows one of three measures of damages: the patentee's lost profits; the infringer's incremental profits; or a 'reasonable royalty' (conceived as the outcome of a hypothetical licensing negotiation (Weil et al. 2001)). Typically, parties employ discounted cash flow methods and 'comparable' licence transactions to support valuation claims. Both *ex ante* and *ex post* methods are used, not always consistently. The law also allows limited consideration of an infringer's *ex ante* alternatives to infringement, such as inventing a substitute. Generally, the most difficult legal and empirical question is: What fraction of (actual or expected) profits should be imputed to the patent? While much damages jurisprudence remains economically ad hoc, courts are increasingly inclined to require the same market analyses that characterize antitrust law (*Crystal Semiconductor v. TriTech Microelectronics*, 246 F. 3d 1336, (Fed. Cir. 2001)).

See Also

- ▶ Patents

Bibliography

Eaton, J., and S. Kortum. 1996. Trade in ideas: Patenting and productivity in the OECD. *Journal of International Economics* 40: 251–278.

Griliches, Z. 1990. Patent statistics as economic indicators: A survey. *Journal of Economic Literature* 28: 1661–1707.

Hall, B., A. Jaffe, and M. Trajtenberg. 2005. Market value and patent citations. *RAND Journal of Economics* 36: 16–38.

Harhoff, D., F. Narin, F.M. Scherer, and K. Vopel. 1999. Citation frequency and the value of patented inventions. *Review of Economics and Statistics* 81: 511–515.

Hausman, J., B. Hall, and Z. Griliches. 1986. Patents and R&D: Is there a lag? *International Economic Review* 27: 265–283.

Henderson, R., A. Jaffe, and M. Trajtenberg. 1998. Universities as a source of commercial technology: A detailed analysis of university patenting, 1965–1988. *Review of Economics and Statistics* 80: 119–127.

Jaffe, A., and M. Trajtenberg. 2002. *Patents, Citations and Innovations: A Window on the Knowledge Economy.* Cambridge, MA: MIT Press.

Jaffe, A., M. Trajtenberg, and R. Henderson. 1993. Geographic localization of knowledge spillovers as evidenced by patent citations. *Quarterly Journal of Economics* 43: 578–598.

Lanjouw, J.O. 1998. Patent protection in the shadow of infringement: simulation estimations of patent value. *Review of Economic Studies* 65: 671–710.

Lanjouw, J.O., and M. Schankerman. 2004. Patent quality and research productivity: Measuring innovation with multiple indicators. *Economic Journal* 114: 441–465.

Lanjouw, J.O., A. Pakes, and J. Putnam. 1998. How to count patents and value intellectual property: Uses of patent renewal and application data. *Journal of Industrial Economics* 46: 405–433.

McCalman, P. 2005. Who enjoys 'TRIPs' abroad? An empirical analysis of intellectual property rights in the Uruguay Round. *Canadian Journal of Economics* 38: 574–603.

Pakes, A. 1986. Patents as options: Some estimates of the value of holding European patent stocks. *Econometrica* 54: 755–784.

Pakes, A., and M. Schankerman. 1984. The rate of obsolescence of patents, research gestation lags, and the private rate of return to research resources. In *R&D, patents and productivity*, ed. Z. Griliches. Chicago: University of Chicago Press.

Pakes, A., and M. Simpson. 1989. Patent renewal data. *Brookings Papers on Economic Acivity: Microeconomics* 1989: 331–410.

Putnam, J. 1996. *The value of international patent rights*. Ph.D. thesis, Yale University.

Schankerman, M. 1998. How valuable is patent protection? Estimates by technology field. *RAND Journal of Economics* 29: 77–107.

Schankerman, M., and A. Pakes. 1986. Estimates of the value of patent rights in European countries during the post-1950 period. *Economic Journal* 96: 1052–1076.

Sullivan, R. 1994. Estimates of the value of patent rights in Great Britain and Ireland, 1852–1876. *Economica* 61: 37–58.

Trajtenberg, M. 1990. A penny for your quotes: Patent citations and the value of innovation. *RAND Journal of Economics* 21: 172–187.

Trajtenberg, M., R. Henderson, and A. Jaffe. 1997. University versus corporate patents: A window on the basicness of invention. *Economics of Innovation and New Technology* 5: 19–50.

Weil, R.L., M.J. Wagner, and P.B. Frank. 2001. *Litigation Services Handbook: The Role of the Financial Expert*. New York: Wiley.

Patents

Bronwyn H. Hall

Abstract

A patent is the legal right of an inventor to exclude others from making or using a particular invention. This right is sometimes termed an 'intellectual property right' and is viewed as an encouragement for innovation. This article gives a brief history of patenting, and discusses the legal and administrative process for obtaining a patent in the major world jurisdictions. Evidence on patent effectiveness in encouraging innovation is surveyed, and the article concludes with a discussion of the use of patent data in economic analysis.

Keywords

Biotechnology industry; Innovation; Intellectual property rights; Inventions; Patent citations; Patent Cooperation Treaty; Patent litigation; Patent races; Patent valuation; Patents; Pharmaceutical industry; Research and development; Trade-Related Aspects of Intellectual Property Rights (TRIPS); World Intellectual Property Organization

JEL Classifications

O3

A patent is the legal right of an inventor to exclude others from making or using a particular invention. This right is customarily limited in time, to 20 years from the date of the application submission in most countries. The principle behind the modern patent is that an inventor is allowed a limited amount of time to exclude others from supplying or using an invention in order to encourage inventive activity by preventing immediate imitation. In return, the inventor is required to make the description and implementation of the invention public rather than keeping it secret, allowing others to build more easily on the knowledge contained in his invention.

The economics of patents has two distinct components, one normative and one positive. The first is directed towards questions of optimal patent policy, the existence and strength of patents, and the design of the patent system. The second uses patent data as an indicator of inventive activity, relying on the fact that patent offices attempt to apply fairly uniform standards of novelty and inventive step when granting patents, so that counts based on them should reflect the innovative activity in a society or in a particular industrial or technology sector. The advantage of patent data is that they are available in great detail over a wide range of time periods, geographic areas, and technological sectors (Griliches 1990). Nevertheless, all patents are not equal, and it is important to understand the operation of patent systems throughout their history in order to make effective use of these data.

This article begins with a brief history of patents, followed by a discussion of the legal and administrative processes for obtaining a patent in the three major patent offices, the United States, European, and Japanese. Then the evidence on patent effectiveness in encouraging innovation is surveyed. The final section discusses the use of patent data in economic analysis.

Brief History

Patents have a long history, although some of the earliest patents are simply the grant of a legal monopoly in a particular good rather than protection of an invention from imitation. Early examples of technology-related patents are Brunelleschi's patent on a boat designed to carry marble up the Arno, issued in Florence in 1421, the Venetian patent law of 1474, and various patent monopolies granted by the English crown between the 15th and 17th centuries. The modern patent, which requires a working model or written description of an invention, dates from the 18th century, first in Britain (1718) and then in the United States (1790), followed closely by France (in both the latter two cases one of the consequences of a revolution). Many other Continental European countries introduced patents during the 19th century, as did Japan. During the 20th century, the use of patent systems became almost universal.

The French patent law of 1791 emphasizes the property right aspect of the patent rather than its use in promoting the useful arts: 'All new discoveries are the property of the author; to assure the inventor the property and temporary enjoyment of his discovery, there shall be delivered to him a patent for five, ten or fifteen years' (Ladas and Parry 2003). In contrast, the Japanese law of 1959 states that its goal is to encourage 'inventions by promoting their protection and utilization and thereby to contribute to the development of industry' (JPO 2006). Patents are enshrined in the US constitution with the sentence 'Congress shall have power . . . to promote the progress of science and useful arts by securing for limited times to authors and inventors the exclusive right to their respective writings and discoveries' (Article 1, Section 8, clause 8), which implicitly recognizes both goals of a patent system, namely, reward to the inventor and the promotion of inventive progress.

In 1883 the Paris Convention for the Protection of Industrial Property ensured national treatment of patent applicants from any country that was a party to it. Its most important provision gave applicants who were nationals or residents of one member state the right to file an application in their own country and then, as long as an application was filed in another country that was a member of the treaty within a specified time (now 12 months) to have the date of filing in the home country count as the effective filing date in that other country (the 'priority date'). This is an important feature of the patent system, and enables worldwide priority to be obtained for an invention originating in any one country, in addition to ensuring that in principle all inventors are treated equally by the system, regardless of the country from which they come.

Legal and Administrative

Although the process for granting a patent varies slightly according to the jurisdiction for which protection is desired, the adoption of the agreement on Trade-Related Aspects of Intellectual Property Rights (TRIPS) in 1995 ensures that it is approximately the same everywhere in the world. This agreement requires its member countries to make patent protection available for any product or process invention in any field of technology with only a few specified exceptions. It also requires them to make the term of protection available for not less than a period of 20 years from the date of filing the patent application.

The World Intellectual Property Organization (WIPO) has almost 200 member states and lists an equivalent number of national patent offices and industrial property offices on its website. In general, the patent right extends only within the border of the jurisdiction that has granted it (usually but not always a country). An important exception is the European system, where it is possible to file a patent application at the European Patent Office (EPO) that will become a set of national patent rights in several European countries at the time of issue (EPO 2006). A similar situation exists with respect to the African Regional Intellectual Property Organization (ARIPO). The exact number and choice of countries is under control of the applicant. Patents granted by the EPO have the same legal status as patents granted by the various national offices that are party to the European Patent Convention (EPC).

The Patent Cooperation Treaty (PCT) came into existence in 1978, and now has 133 countries as contracting signatories. Any resident or national of a contracting state of the PCT may file an international application under the PCT that specifies the office which should conduct the search. The PCT application serves as an application filed in each designated contracting state. However, in order to obtain patent protection in a particular state, a patent needs to be granted by that state to the claimed invention contained in the international application. The advantage of a PCT application is that fewer searches need to be conducted and the process is therefore less expensive. In fact, 87 per cent of the PCT applications go to one of three patent office for search: those in the United States, Europe, and Japan. Most of the other systems rely on them for the search process and follow them in a number of other areas. Therefore the brief account that follows focuses on these three major systems.

EPO patent grants are issued for inventions that are novel, mark an inventive step, are commercially applicable, and are not excluded from patentability for other reasons (Article 52, EPC). The statutory requirements for patentability in the United States are similar: ‘any new and useful process, machine, manufacture, or composition of matter, or any new and useful improvement thereof’ may be patented (35 US Code 101-103 and 112). By itself, this definition does not create a subject matter restriction, although it has long been held that laws of nature, physical phenomena, and abstract ideas are not patentable subject matter.

The origins of the Japanese patent system date back to the Meiji Era (1868–1912). Early patent laws in 1885 and 1899 were modeled on French, US, and then German patent law. In 1899, Japan acceded to the Paris Convention for the protection of industrial property. The patent law was completely revised in 1909, 1921, and 1959. Today, in Japan, patent rights are still protected by the Patent Act of 1959, frequently amended since then (JPO 2006; Kotabe 1992). Two important recent changes were the introduction of a product patent in 1976 and the switch to allowing multiple claims in a patent in 1987, both of which have the effect of bringing the system closer to those in Europe and the United States (Nagaoka 2006).

US patent applications must be filed within one year of the invention’s public use or publication – this year is called the ‘grace period’, intended to allow researchers some ability to publish their results as soon as possible. In Europe and other jurisdictions, there is no grace period. Alone among the world’s patent offices, the US Patent and Trademark Office operates a ‘first-to-invent’ rather than a ‘first inventor-to-file’ system. In either case, the applicant must be the inventor (except in certain special cases such as death or mental incapacity), but in the US system priority is assigned to the inventor who can show that he reduced the invention to practice first. Also unique to the United States is the fact that patent applications are not made public automatically. Ordinarily patent applications are published 18 months after their priority date, but in the United States an applicant may request exemption from this rule if he files an application on the equivalent invention only at the United States Patent and Trademark Office (USPTO) and in no other jurisdiction.

Many patent offices have a provision for challenging patents following their issue. In the United States, any third party may request re-examination of a patent during its lifetime, although for various reasons related to potential subsequent litigation this opportunity is rarely taken up. In Europe and Japan, robust patent opposition systems with limited time frames operate, and these systems are often used by rival firms as an alternative to more expensive litigation (Hall et al. 2003). In Europe this avenue of challenge is particularly attractive because it is the last opportunity to attack a patent at the European-wide level rather than in individual national courts.

Patents are valuable only if they can be enforced and this fact has a number of implications for their use. First, the ability of the courts to reach the ‘correct’ verdict with respect to infringement and validity will matter; in situations or jurisdictions where there is a great deal of uncertainty about the outcome, and even if both parties agree as to the merits of the case, it may be worth pursuing the issue further or in some cases, reaching a private financial settlement to avoid a random outcome in the courts. Second, the costs of litigation will matter: parties with deep pockets can threaten those with less access to resources, or where the opportunity cost of paying attention to a patent suit is high. On the other hand smaller parties with less to lose can also hold up firms with large sunk investments that they might lose. Finally, the threat of litigation may discourage firms from even entering certain areas, thus providing a disincentive rather than an incentive for R&D. Lerner (1995) documented this phenomenon for biotechnology. The degree to which these kinds of threats matter depends to a great extent on the costs and extent of litigation, both of which tend to be higher in the United States than in many other countries.

Research on patent litigation is difficult because of the data collection problem (it frequently requires accessing the records of courts in several different jurisdictions) but in recent years there have been series of studies of US patent litigation (Moore 2000; Lanjouw and Schankerman 2001; Bessen and Meurer 2005) and at least one of the German system (Cremers 2004). All of these studies document the fact that litigated patents tend to be the more valuable. The US studies also show that only about five per cent of such suits go to trial, with the remainder being settled before going to trial. They also show that whether patent litigation has increased depends on whether it is measured in aggregate or per patent. That is, the increase in patent litigation has roughly paralleled the increase in patenting, at least in the United States.

Economics of Patents

The economic view of patents is that they offer a bargain between society and the inventor: in return for a limited period of exclusivity, the inventor agrees to make his invention public rather than keeping it secret. Therefore, one of the central questions that arises when patents are used as a policy tool to encourage innovation is whether this tool is effective. The theoretical literature in this area produces somewhat ambiguous results. In the simplest case, where a patent corresponds to a single product and knowledge is not cumulative, clearly patents do encourage innovation. In fact, the early theoretical industrial organization literature on patent races seemed to suggest that patents produced too much innovation (Wright 1983; Reinganum 1989). However, models that incorporate the cumulative nature of innovation or the fact that production of something new frequently relies on patents held by a large number of entities produce more ambiguous results (Judd 1985; Bessen and Maskin 2006).

This question has also proved exceedingly difficult to answer empirically, largely because of the absence of real experiments. Some researchers have looked at historical eras when there were changes to the system and examined the consequences for subsequent innovative activity, measured either by patenting in a jurisdiction not affected by the changes to the system or by invention counts obtained independently (Lerner 2002; Moser 2005). A second widely used approach is to survey firms and ask about their patent use (Levin et al. 1987; Cohen et al. 2002; Arundel 2003). Using these kinds of survey data matched to R&D spending and innovation outcomes, more structural approaches have been pursued by Baldwin et al. 2000; Arora et al. 2003; and Bloom et al. 2005, among others.

A few conclusions emerge from this body of work. First, introducing or strengthening a patent system (lengthening the patent term, broadening subject matter coverage or available scope, improving enforcement) unambiguously results in an increase in patenting and also in use of patents as a tool of firm strategy (Lerner 2002; Hall and Ziedonis 2001). It is much less clear that these changes result in an increase in innovative activity, although they may redirect such activity toward things that are patentable and are not subject to being kept secret within the firm (Moser 2005). Sakakibara and Branstetter (2001) studied the effects of expanding patent scope in Japan in 1988 and found that this change to the patent system had a very small effect on R&D activity in Japanese firms.

The survey evidence from a number of countries shows rather conclusively that patents are not among the important means to appropriate returns to innovation, except perhaps in pharmaceuticals (Levin et al. 1987; Cohen et al. 2002; Arundel 2003). More important means of appropriation are usually superior sales and service, lead time, and secrecy. Patents are usually rated as important only for blocking and defensive purposes. Thus, if there is an increase in innovation due to patents, it is likely to be centred in the pharmaceutical and biotechnology areas, and possibly specialty chemicals. Arora et al. (2003) found that increasing the patent premium, which they describe as the difference in payoffs to patented and unpatented inventions, does not increase R&D much except in pharmaceuticals and biotechnology. Using aggregate data across 60 countries for the 1960–90 period, Ginarte and Park (1997) found that the strength of the patent system is positively associated with R&D investment in countries with high median incomes (that is, G-7 and others), but not in lower-income countries.

Recently it has been suggested that the existence and strength of the patent system affects the organization of industry by allowing trade in knowledge, which facilitates the vertical disintegration of knowledge-based industries and the entry of new firms that possess only intangible assets. The argument is that, by creating a strong property right for the intangible asset, the patent system enables activities that formerly had to be kept within the firm because of secrecy and contracting problems to move out into separate entities. Although limited, research in this area supports this conclusion in the chemical and semiconductor industries (Arora et al. 2001; Hall and Ziedonis 2001).

Economic analysis has also been used to address the optimal design of the patent system. The seminal work in this area was Nordhaus (1969), which considered two policy instruments: the length of the patent term and the breadth of the patent, that is, the range or scope of the inventions covered. The broader the scope of a patent, the larger the number of competing products and processes that will infringe the patent, and the larger the market power of the patentholder. Later work by Gilbert and Shapiro (1990) and Klemperer (1990) built on and extend his method of analysis. Unfortunately, even though all three sets of authors simplified the problem by assuming that a patent corresponds to a product and that there is no uncertainty, the welfare conclusions still turn on assumptions about the nature of the product market and the existence of close substitutes for the patented product. The main conclusion from this line of work is that optimal patent design is likely to depend on the nature of the product market and the technology, which is inconsistent with long-standing practice and policy in most patent systems. Historically, the only important exception to the homogeneous treatment of technologies is the extreme one of excluding some of them (such as pharmaceutical products, medical practices, or disembodied software) completely from the system.

Recent theoretical and empirical work on the patent system has focused on a set of questions that have increased in importance because of the complexity of modern technology and the growth in patent use in sectors that traditionally had paid relatively little attention to them. Briefly described, the new setting is one where a single product involves hundreds of patents, and where one innovation builds directly on many others. Neither feature is really new, but both have assumed increasing importance in a number of technology areas such as information technology and biotechnology. At a theoretical level, Scotchmer (1991, 2005) was the first to identify the problem that cumulative innovation creates for the patent system, in the sense that it is difficult if not impossible to set incentives at the correct level for both the first and subsequent innovators.

When development of an innovative product requires multiple patent inputs, Heller and Eisenberg (1998) have argued forcefully that the licensing solution may fail because of transactions costs if a large number of patentholders are involved. One consequence of this fragmentation threat may be increased defensive patenting by the product developer. Empirical evidence for this proposition has been provided by Ziedonis (2004) in the context of the semiconductor industry.

Using Patent Data

Researchers into the economics of innovation and technical change frequently find themselves in need of measures of innovative output or success, preferably classified by sector or technology. Many would also like measures of knowledge flow between individuals and firms, given the potential importance of spillovers in the production of knowledge. In recent years, the growth in importance of the knowledge economy worldwide has lead to an increased interest in such measures. As was noted long ago by such pioneers in the field as Schmookler (1966), patent data can be very helpful in constructing them. The primary advantage of patent data is that they are available over a wide range of countries and years, for detailed technology classes, and they contain information on inventor, geographic area, and owner (if there is one other than the inventor). Together, these data provide information on the locus and type of newly created knowledge. The second advantage is that they provide information on links between different quanta of knowledge via the citations to other patents and non-patent documents that they contain (see Jaffe et al. 2000, for further justification of the use of patent citations to model knowledge flow and for the limitations of the measure). With the possible exception of data on scientific paper publication, no other data source comes even close to providing this level and quantity of information about the creation and dissemination of new knowledge.

The use of patent data as a proxy for innovation output in the economic analysis of technological change dates back to the path-breaking analyses of Schmookler (1966) and Scherer (1965). An overview is given in OECD (1994). The availability of information from the US patent office in machine-readable form in the late 1970s enabled research using these data with much larger samples of firms; the resulting early work is reported in Griliches (1984) and then surveyed by Griliches et al. (1987) and Griliches (1990). At the same time, Schankerman and Pakes (1986) pioneered the use of renewal data from the patent offices of several European countries to estimate the value distribution of patents; at the time, such data were not available for the United States owing to the absence of renewal fees in that country.

The results of this early work were, first, to demonstrate a strong correlation between the size of a firm's R&D effort and its patenting output, with little evidence that smaller programmes and firms yielded more output per unit of input, once selection was controlled for. Second, the renewal data, along with pieces of evidence from some specific sectors such as pharmaceuticals (Grabowski and Vernon 1994) and medical devices (Trajtenberg 1990), suggested that the value distribution of patents was very skewed, with a few patents worth a lot and most patents worth nothing. Third, there was little evidence that patent outcomes added much predictive power to sales, profits, or market value equations in the presence of R&D expenditure (Griliches et al. 1991).

With the advent of the personal computer and the increased access to computing power on the part of economic researchers, it became feasible to construct data-sets containing patent citations in the late 1980s, leading to a second wave of research. Similarly to a research paper, the patent document contains a set of references to earlier patents and scientific literature on which it builds; a typical patent referenced approximate five earlier patents in the 1980s, and an increasing number as time passes. These citations can be used to give an indication of the impact of a patented invention on the inventions in subsequent patents and to investigate an additional set of questions related to the flow of knowledge across time, space and organizational boundaries. However, it is important to note that differences exist in citation practice between the US and other patent systems (see Webb et al. 2005; and Harhoff et al. 2006, for further discussion of this issue), and most of the validation of this methodology has been done using US data.

Researchers have used these data to explore questions involving spatial spillovers (for example, Jaffe et al. 1993), knowledge flows among firms in a research consortium (for example, Ziedonis et al. 1998), and spillovers from public research (for example, Jaffe and Trajtenberg 1996; Jaffe and Lerner 2001). In using citations as evidence of spillovers, or at least knowledge flows, from cited inventors to citing inventors, it is clearly a problem that many of the citations are added by the inventor's patent attorney or the patent examiner, and may represent inventions that were wholly unknown to the citing inventor. On the other hand, in using citations received by a patent as an indication of that patent's importance, impact or even economic value, the citations that are identified by parties other than the citing inventor may well convey valuable information about the size of the technological 'footprint' of the cited patent.

Beginning with Trajtenberg's (1990) study of the welfare impact of CAT scanners, there are by now a number of studies that 'validate' the use of citations data to measure economic impact, by showing that citations are correlated with non-patent-based measures of value. Hall et al. (2005) investigated the use of citations as an indicator of private invention value in a large sample of publicly traded US manufacturing firms and confirmed that, although patent yield conveys little information beyond that conveyed by R&D spending, citation-weighted patents are strongly related to market value in a nonlinear way, with very highly cited patents worth a great deal more than those with less than average citation.

Recent work by Lanjouw and Schankerman (2004) also uses citations, together with other attributes of the patent (number of claims and number of different countries in which an invention is patented) as a proxy for patent quality. They find that a patent 'quality' measure based on these multiple indicators has significant power in predicting which patents will be renewed and which will be litigated. They infer from this that these quality measures are significantly associated with the private value of patents. Similarly, Harhoff et al. (1999) surveyed 962 holders of German patents that had a priority date of 1977, asking them to estimate at what price they would have been willing to sell the patent right in 1980, about three years after the date at which the German patent was filed. They find both that more valuable patents are more likely to be renewed to full term and that the estimated value is correlated with subsequent citations to that patent. As in Hall et al. (2005, p. 23), the most highly cited patents are very valuable, 'with a single U.S. citation implying on average more than $1 million of economic value'.

See Also

▶ Intellectual Property

Bibliography

Arora, A., A. Fosfuri, and A. Gambardella. 2001. *Markets for technology: The economics of innovation and corporate strategy*. Cambridge, MA: MIT Press.

Arora, A., M. Ceccagnoli, and W. Cohen. 2003. *R&D and the patent premium*, Working paper, no. 9431. Cambridge, MA: NBER.

Arundel, A. 2003. *Patents in the knowledge-based economy: Report of the know survey.* Maastricht: MERIT, University of Maastricht.

Baldwin, J.R., P. Hanl, and D. Sabourin. 2000. *Determinants of innovative activity in Canadian manufacturing firms: The role of intellectual property rights*, Working paper, no. 122. Ottawa: Statistics Canada.

Bessen, J., and E. Maskin. 2006. *Sequential innovation, patents, and imitation*, Working paper. Boston University School of Law and Princeton University.

Bessen, J., and M. Meurer. 2005. *The patent litigation explosion. Law and economics*,Working paper, no. 05-18. Boston University School of Law.

Bloom, N., J. Van Reenen, and M. Schankerman. 2005. *Identifying technology spillovers and product market rivalry*, Discussion paper, no. 3916. London: CEPR.

Cohen, W.M., A. Goto, A. Nagata, R.R. Nelson, and J.P. Walsh. 2002. R&D spillovers, patents and the incentives to innovate in Japan and the United States. *Research Policy* 31: 1349–1367.

Cremers, K. 2004. *Determinants of patent litigation in Germany*, Discussion paper, no. 04-72. Mannheim: Centre for European Economic Research (ZEW).

EPO (European Patent Office). 2006. Online. Available at http://www.european-patent- office.org/index.en.php. Accessed 4 Jan 2007.

Federal Trade Commission (FTC). 2003. *To promote innovation – The proper balance of competition and patent law and policy.* Washington, DC: FTC.

Gilbert, R., and C. Shapiro. 1990. Optimal patent length and breadth. *RAND Journal of Economics* 21: 106–112.

Ginarte, J.C., and W. Park. 1997. Determinants of patent rights: A cross-national study. *Research Policy* 26: 283–301.

Grabowski, H., and J.M. Vernon. 1994. Returns to R&D on new drug introductions in the 1980s. *Journal of Health Economics* 13: 383–406.

Griliches, Z., ed. 1984. *R&D, patents and productivity.* Chicago: University of Chicago Press.

Griliches, Z. 1990. Patent statistics as economic indicators: A survey. *Journal of Economic Literature* 28: 1661–1707.

Griliches, Z., A. Pakes, and B.H. Hall. 1987. The value of patents as indicators of inventive activity. In *Economic policy and technological performance*, ed. P. Dasgupta and P. Stoneman. Cambridge: Cambridge University Press.

Griliches, Z., B.H. Hall, and A. Pakes. 1991. R&D, patents, and market value revisited: Is there a second (technological opportunity) factor? *Economics of Innovation and New Technology* 1 (3): 183–202.

Hall, B.H., and R. Ziedonis. 2001. The patent paradox revisited: An empirical study of patenting in the US semiconductor industry, 1979–1995. *RAND Journal of Economics* 32: 101–128.

Hall, B.H., S. Graham, D. Harhoff, and D.C. Mowery. 2003. Prospects for improving U.S. patent quality via postgrant opposition. *Innovation Policy and the Economy* 4: 115–143.

Hall, B.H., A.B.. Jaffe, and M. Trajtenberg. 2005. Market value and patent citations. *RAND Journal of Economics* 36: 16–38.

Harhoff, D., F. Narin, F.M. Scherer, and K. Vopel. 1999. Citation frequency and the value of patented inventions. *Review of Economics and Statistics* 81: 511–515.

Harhoff, D., K. Hoisl, and C. Webb. 2006. *European patent citations – How to count and how to interpret them?* Working paper. University of Munich, CEPR and OECD.

Heller, M.A., and R.S. Eisenberg. 1998. Can patents deter innovation? The anticommons in biomedical research. *Science* 280: 698–701.

Jaffe, A.B.., and J. Lerner. 2001. Reinventing public R&D: Patent policy and the commercialization of national laboratory technologies. *RAND Journal of Economics* 32: 167–198.

Jaffe, A.B.., and M. Trajtenberg. 1996. Flows of knowledge from universities and federal labs: Modeling the flow of patent citations over time and across institutional and geographic boundaries. *Proceedings of the National Academy of Sciences* 93: 12671–12677.

Jaffe, A.B.., M. Trajtenberg, and R. Henderson. 1993. Geographic localization of knowledge spillovers as evidenced by patent citations. *Quarterly Journal of Economics* 108: 577–598.

Jaffe, A.B.., M. Trajtenberg, and M. Fogarty. 2000. *The meaning of patent citations: Report of the NBER/Case Western Reserve survey of patentees*, Working paper no. 763. Cambridge, MA: NBER.

JPO (Japanese Patent Office). 2006. A history of system of industrial property rights. Online. Available at http://www.deux.jpo.go.jp/cgi/search.cgi?query = history&lang = en&root = short. Accessed 19 Dec 2006.

Judd, K.L. 1985. On the performance of patents. *Econometrica* 53: 567–595.

Klemperer, P. 1990. How broad should the scope of patent protection be? *RAND Journal of Economics* 21: 113–130.

Kotabe, M. 1992. A comparative study of U.S. and Japanese patent systems. *Journal of International Business Studies* 23: 147–168.

Ladas and Parry LLP. 2003. A brief history of the patent law of the United States. Online. Available at http://www.ladas.com/Patents/USPatentHistory.html. Accessed 19 Dec 2006.

Lanjouw, J.O., and M. Schankerman. 2001. Characteristics of patent litigation: A window on competition. *RAND Journal of Economics* 32: 129–151.

Lanjouw, J.O., and M. Schankerman. 2004. Patent quality and research productivity: Measuring innovation with multiple indicators. *Economic Journal* 114: 441–465.

Lerner, J. 1995. Patenting in the shadow of competitors. *Journal of Law and Economics* 38: 463–495.

Lerner, J. 2002. Patent policy shifts and innovation over 150 years. *American Economic Review* 92: 221–225.

Levin, R.C., A.K. Klevorick, R.R. Nelson, and S.G. Winter. 1987. Appropriating the returns from industrial research and development. *Brookings Papers on Economic Activity* 1987 (3): 783–831.

Moore, K.A. 2000. Judges, Juries, and patent cases – An empirical peek inside the black box. *Michigan Law Review* 99 (281): 365–409.

Moser, P. 2005. How do patent laws influence innovation? Evidence from nineteenth-century world fairs. *American Economic Review* 95: 1214–1236.

Nagaoka, S. 2006. Reform of patent system in Japan and challenges. Paper presented at the conference on 21st century innovation systems for Japan and the United States: Lessons from a decade of change. Tokyo: Institute of Innovation Research, Hitotsubashi University.

Nordhaus, W. 1969. *Invention, growth and welfare: A theoretical treatment of technological change*. Cambridge, MA: MIT Press.

OECD. 1994. *The measurement of scientific and technological activities: Using patent data as science and technology indicators*. Paris: OECD.

Reinganum, J.F. 1989. The timing of innovation: Research, development, and diffusion. In *Handbook of industrial organization*, ed. R. Schmalensee and R.D. Willig, vol. 1. Amsterdam: North-Holland.

Sakakibara, M., and L. Branstetter. 2001. Do stronger patents induce more innovation? Evidence from the 1988 Japanese patent law reforms. *RAND Journal of Economics* 32: 77–100.

Schankerman, M., and A. Pakes. 1986. Estimates of the value of patent rights in European countries during the post-1950 period. *Economic Journal* 96: 1052–1076.

Scherer, F.M. 1965. Firm size, market structure, opportunity, and the output of patented innovations. *American Economic Review* 55: 1097–1123.

Schmookler, J. 1966. *Invention and economic growth*. Cambridge, MA: Harvard University Press.

Scotchmer, S. 1991. Standing on the shoulders of giants. *Journal of Economic Perspectives* 5 (1): 29–41.

Scotchmer, S. 2005. *Innovation and incentives*. Cambridge, MA: MIT Press.

Trajtenberg, M. 1990. A penny for your quotes: Patent citation and the value of innovations. *RAND Journal of Economics* 21: 172–187.

USPTO (United States Patent and Trademark Office). Online. Available at https://www.uspto.gov. Accessed 4 Jan 2007.

Webb, C., H. Dernis, D. Harhoff, and K. Hoisl. 2005. *Analysing European and international patent citations: A set of EPO patent database building blocks*, STI working paper, no. 2005/9. Paris: OECD.

WIPO (World Intellectual Property Organization). Online. Available at https://www.wipo.int/patentscope/en/. Accessed 4 Jan 2007.

Wright, B.D. 1983. The economics of invention incentives: Patents, prizes, and research contracts. *American Economic Review* 73: 691–707.

Ziedonis, R.H. 2004. Don't fence me in: Fragmented markets for technology and the patent acquisition strategies of firms. *Management Science* 50: 804–820.

Ziedonis, A.A., R.H. Ziedonis, and B.S. Silverman. 1998. *Research consortia and the diffusion of technological knowledge: Insights from SEMATECH*, Working paper. University of Michigan and University of Toronto.

Path Analysis

İnsan Tunali

Path analysis is a method for estimating and testing the internal consistency of models with a postulated causal structure. The postulated structure is displayed in the form of path diagrams, where one-way arrows link causal variables to their outcomes, and curved two-headed arrows connect related variables whose causal links are not under study. Estimation proceeds along the lines of method of moments and instrumental variables theory: the causal ordering of variables along distinct paths are exploited to express the unknown structural parameters in terms of the population moments of the observed and the unobserved variables. Estimating equations are obtained by replacing the population moments of the observed variables by their sample counterparts, which are then solved for the unknown parameters and the estimates of the moments of the unobservables (which themselves can be thought of as structural parameters).

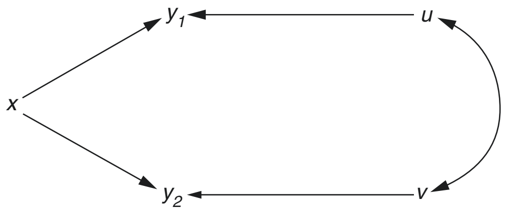
Consider the following simple example discussed in Wright (1960) and Duncan (1975). The structural model consists of

$$y_1 = \beta_1 x + u$$

Model (1)

$$y_2 = \beta_2 x + \upsilon$$

where y's denote the observed (endogenous) variables, x the observed (exogenous) variable, u and υ the unobserved (exogenous) variables or disturbances, and β's the unknown parameters. To keep the algebra simple, all exogenous variables are assumed to have zero means here and below. The disturbances are assumed to have non-zero variances and covariance, and are uncorrelated with x. The path diagram depicting the postulated causal structure is given in Fig. 1. The



Path Analysis, Fig. 1 The path diagram for Model (1)

relationships between the population moments and the structural parameters are easily derived to be

$$\sigma_{11}=\beta_1^2\sigma_{xx}+\sigma_{uu},\sigma_{12}=\beta_1\beta_2\sigma_{xx}+\sigma_{u\upsilon},\sigma_{1x}=\beta_1\sigma_{xx}$$
$$\sigma 22=\beta_2^2\sigma_{xx}+\sigma_{\upsilon\upsilon},\sigma_{2x}=\beta_2\sigma_{xx} \qquad (1')$$

where the variance–covariance structures of the observables y_1, y_2, x and the unobservables u, υ are respectively given by

$$\begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{1x} \\ & \sigma_{22} & \sigma_{2x} \\ & & \sigma_{xx} \end{bmatrix}, \quad \begin{bmatrix} \sigma_{uu} & \sigma_{u\upsilon} \\ & \sigma_{\upsilon\upsilon} \end{bmatrix}.$$

Inspection reveals that the five equations in (1′) will uniquely determine the five unknowns, β_1, β_2, σ_{uu}, $\sigma_{\upsilon\upsilon}$, $\sigma_{u\upsilon}$, using estimates obtained from a random sample on the observables to replace σ_{11}, σ_{22}, σ_{12}, σ_{xx}, σ_{1x} and σ_{2x}.

To see how path analysis can be utilized to test the internal consistency of a postulated structure, consider the same set of equations as in model (1), but omit the curved arrow from Fig. 1. This is equivalent to setting $\sigma_{u\upsilon} = 0$ in the equation system (1′). The five equations now over-determine the four unknowns, β_1, β_2, σ_{uu} and $\sigma_{\upsilon\upsilon}$. Simple algebraic manipulation of the covariance terms reveals that the solution to the system will be unique if and only if $\sigma_{12} = \sigma_{1x}\,\sigma_{2x}/\sigma_{xx}$. This condition (referred to as an 'over-identifying restriction' in the econometrics literature) can be tested using the sample counterparts of the population moments involved. Note that no such check on the internal consistency of the original model is available: Model (1) is 'just-identified'.

The preceding example illustrates the basic ideas behind path analysis, cast within the conventional linear regression framework. The method's origins, however, lie elsewhere. Path analysis was invented by the geneticist Sewall Wright, whose work in the 1920s foreshadowed the econometric literature on structural estimation. Wright formulated complex models with unobservables and wrestled with simultaneity and identification long before econometricians began their systematic study of these topics. Wright's main subject of study, heritability, offered him the necessary insights for modelling the links between cause and effect. His objective was to infer correlations between the traits of interest – bone sizes of rabbits, skin colour and birth weight of guinea pigs, etc. – across different generations in a population. Towards this end, Wright devised an algorithm for reading off the estimating equations directly from the path diagram. For a model which does not depict simultaneity, Wright described his algorithm as follows (cf. Wright 1960):

> The correlation between any two variables in a properly constructed diagram of relations is equal to the sum of contributions pertaining to the paths by which one may trace from one to the other in the diagram without going back after going forward along an arrow and without passing through any variable twice in the same path. A coefficient pertaining to the whole path connecting two variables, and thus measuring the contribution of that path to the correlation, is known as a *compound path coefficient*. Its value is the product of the coefficients pertaining to the elementary paths along its course. One, but not more than one of these, may pertain to a two-headed arrow without violating the rule against going back after going forward.

Since the analysis of correlations constituted his primary interest, Wright worked with standardized variables, having zero means and unit variances. His algorithm applied to the model above without standardization of the variances would yield the system in (1′). If standardized variables were utilized instead, the resulting equations would be in terms of simple correlations and beta coefficients, referred to as 'path coefficients' by Wright. It is a matter of simple algebra to convert the above equations to Wright's equations, and vice versa. For example, taking the first expression in (1′) above and dividing by σ_{11}, we get the equivalent representation.

$$1 = \left[(\beta_1)(\sigma_{xx}/\sigma_{11})^{1/2}\right]^2 + \left[(1)(\sigma_{uu}/\sigma_{11})^{1/2}\right]^2$$
$$= p_{1x}^2 + p_{1u}^2,$$

where p's denote the path coefficients. (Note that the structural coefficient of u has been entered as '1' in this expression.) Wright's representation has the advantage of providing a readily interpretable goodness of fit measure: p_{1x}^2 and p_{1u}^2 and p are respectively the proportion of the variation in y_1 that can be explained by x, and that which is left unexplained.

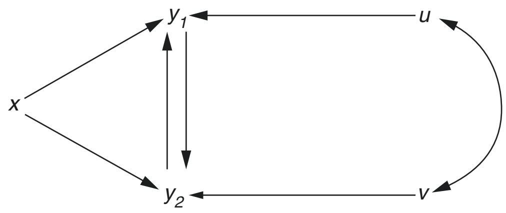
Path analysis is capable of handling a much more general class of problems than the one discussed above. To illustrate the nature of the extensions, we look at some other simple models. These and other examples may be found in Goldberger (1973) and Duncan (1975). We first consider the simultaneous equations model

$$y_1 = \gamma_1 y_2 + \beta_2 x + u$$

Model (2)

$$y_2 = \gamma_2 y_1 + \beta_2 x + \upsilon$$

where u and υ are unobserved, and are assumed to be uncorrelated with x. The associated path diagram is given in Fig. 2. It is straightforward to show that this model is not identified without further assumptions. Clearly, setting $\gamma_1 = \gamma_2 = 0$ gives model (1). Setting $\gamma_1 = \sigma_{u\upsilon} = 0$ (which is equivalent to removing the arrow going from y_2 to y_1 and the double-headed arrow connecting u and υ in Fig. 2) gives a recursive model, which can be shown to be just-identified. Recursive models have been studied extensively in the sociology literature, where Wright's work has had its significant impact (cf. Boudon 1965; Duncan 1966).



Path Analysis, Fig. 2 Path diagram for Model (2)

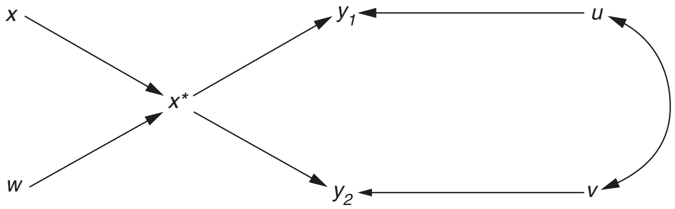
Next, we consider the latent variable model

$$\text{Model}(3) \qquad \begin{aligned} x^* &= \alpha x + w \\ y_1 &= \beta_1 x^* + u \\ y_2 &= \beta_2 x^* + \upsilon \end{aligned}$$

where y's are observed, x^* is unobserved, w is unobserved and uncorrelated with the other disturbances u and υ, and x is observed and uncorrelated with w, u, υ. The path diagram corresponding to this specification is given in Fig. 3.

It can be shown that Model (3) is not identified without further assumptions. Clearly, setting $\alpha = 1$ and $w = 0$ yields Model (1), which is just-identified. The versatility of path analysis can be underscored by nothing other versions of (3) that appear under various names in the literature on linear structural equation models. (We ignore the issue of identification for these models.) Setting $\alpha = 1$ we obtain an errors in variables (or measurement error) model. Dropping the x^* equation from (3), we get a factor analytic model, with x^* as the common factor, and u, υ as the specific factors. If we replace the scalars α and x by conformable vectors α' and x, we obtain a multiple indicator-multiple cause (MIMIC) model. Finally, note that the simultaneous equation model (2) and the latent variable model (3) can be combined to arrive at a more general structural model, widely known by the name of the computer program used for

Path Analysis, Fig. 3 Path diagram for Model (3)



estimating such models, LISREL (cf. Jöreskog and Sörbom 1981).

Bibliographical notes. Blalock (1971) contains Wright (1960) and Duncan (1966), as well as other papers of historical interest. Goldberger (1972) reviews Wright's contributions from the point of view of the econometrics literature on structural estimation. Hauser and Goldberger (1971) establish the links between the path analysis literature and the econometrics and psychometrics literatures. Bentler (1983) provides an overview of the state of the art on linear structural equation models.

See Also

- ▶ Causal Inference
- ▶ Econometrics
- ▶ Opportunity Cost
- ▶ Regression and Correlation Analysis

Bibliography

Bentler, P.M. 1983. Simultaneous equation systems as moment structure models – With an introduction to latent variable models. *Journal of Econometrics* 22: 13–42.

Blalock, H.M. (ed.). 1971. *Causal models in the social sciences*. Chicago: Aldine.

Boudon, R. 1965. A method of linear causal analysis: Dependence analysis. *American Sociological Review* 30: 365–373.

Duncan, O.D. 1966. Path analysis: Sociological examples. *American Journal of Sociology* 72: 1–16.

Duncan, O.D. 1975. *Introduction to structural equation models*. New York: Academic.

Goldberger, A.S. 1972. Structural equation methods in the social sciences. *Econometrica* 40: 979–1001.

Goldberger, A.S. 1973. Structural equation models: An overview. In *Structural equation models in the social sciences*, ed. A.S. Goldberger and O.D. Duncan. New York: Seminar Press.

Hauser, R.M., and A.S. Goldberger. 1971. The treatment of unobservable variables in path analysis. In *Sociological methodology 1971*, ed. H.L. Costner. San Francisco: Jossey-Bass.

Jöreskog, K.G., and D. Sörbom. 1981. *LISREL: Analysis of linear structural relationships by the method of maximum likelihood – User's guide*. Chicago: National Educational Resources.

Wright, S. 1960. Path coefficients and path regressions: Alternative or complementary concepts? *Biometrics* 16: 189–202.

Path Dependence

Steven N. Durlauf

Abstract

Path dependence refers to the idea that 'history matters', that is, that various types of contingent events may have long-term consequences. This article provides some formalization of the concept and assesses its usefulness in elucidating economic phenomena.

Keywords

Ergodicity and non-ergodicity in economics; Multiple equilibria; Multiple steady states; Network externalities; Path dependence; QWERTY keyboard configuration; Reflection problem

JEL Classification

D83; D85; L14; L15; O3; Z13

Originating with work of Paul David (1985, 1986) and W. Brian Arthur (1989), there have been a number of efforts by economists to argue that path dependence exists in various socio-economic outcomes. Heuristically, path dependence is understood to mean that 'history matters' in the sense that certain long-term economic outcomes are contingent on particular events that themselves need not have occurred.

The canonical example of path dependence is that of the QWERTY keyboard configuration for typewriters, studied by David (1985, 1986). David argues that the emergence of the QWERTY configuration as the standard for typewriters was an historical accident in the sense it was the consequence of a set of decentralized, uncoordinated choices by different economic actors whose decisions were driven by network externalities. As such, the standard became locked in even though it was not socially optimal, that is, there existed an alternative configuration, the Dvorak keyboard, which was preferable in terms of typing efficiency. From this perspective, the QWERTY keyboard was simply one of several potential long-run standards that could have emerged, its actual emergence being a function of a particular set of contingent events, that is, shocks. Other cases where path dependence has been argued to occur include nuclear reactor technology (Cowan 1990) and railway gauge length (Puffert 2002). Arthur (1989) provides a formal model of path dependence which captures the logic of the QWERTY example; see also Farrell and Saloner (1985) for an early and important example in the industrial organization literature on network externalities which exhibits path dependence-type phenomena.

From the perspective of current economics, most of the path dependence literature is somewhat anomalous. While path dependence is often informally invoked to describe one phenomenon or another, there has been little systematic research on path dependence outside of economic history; in particular, there is a general dearth of formal theoretical and econometric analyses. As a result, discussions of path dependence are often very imprecise. Such imprecision may first be seen in definitions of path dependence. The term means different things in different writings, so that disagreements on its presence to some extent are simply disagreements about its definition.

Most discussions seem to equate path dependence with non-ergodicity. Consider a set of independent shocks ε_t and an outcome of interest x_t; let μ denote a probability measure. The process x_t is non-ergodic if for a fixed k, $\lim_{j\text{ß¥}} \mu\left(x_{t+j} | \varepsilon_0, \ldots, \varepsilon_k\right)$ depends on the realization of ε_0, …, ε_k. Such a definition captures some of the main intuition underlying qualitative discussions of path dependence in that for such processes history matters, that is, particular sets of shocks have long-run consequences. Theoretical models of path dependence such as Arthur (1989) have this property. It is worth noting that, from the perspective of those few formal theories that claim to model path dependence, the phenomenon is typically a form of multiple equilibria or multiple steady states, both of which occur in many contexts. See Blume and Durlauf (2001) for a conceptual discussion on how various deviations from the Arrow–Debreu baseline, when combined with complementarities (those of network externalities are simply one example), can lead to multiple equilibia or multiple metastable states, that is, states from which a system will emerge only after long epochs. What distinguishes theoretical models of path dependence appears to be the explicit attention to the consequences of individuals making decisions sequentially, so that dynamic forms of coordination failure can occur.

However, it is far from clear that such a notion of path dependence as non-ergodicity is sensible for the examples for which path dependence has been claimed to occur. While it is incontrovertible that technological standards are subject to strong network externalities, it is equally true that technological standards evolve over time. One early example of path dependence was the success of the VHS tape standard over Betamax. In light of the rise of the DVD, it is unclear in what sense the success of the VHS tape over a particular time horizon is evidence of anything deeper than network externalities per se. It is possible that a better definition of path dependence relates to whether shocks to a system are self-reversing. Suppose we consider a system where $\varepsilon_l = 0, \quad l > k$. If $\lim_{j\text{ß¥}} \mu\left(x_{t+j} | \varepsilon_0, \ldots, \varepsilon_k, \;\; \varepsilon_l = 0, \;\; l < k\right)$, depends on the realization of ε_0, …, ε_k, then one has a system in which shocks to a system can persist unless overcome by future shocks. This notion of path dependence may be more sensible for contexts such as technological standards as it respects the role of new technologies in undoing current configurations; in fact, it seems the more appropriate definition from the perspective of various examples in economic history. This definition of path dependence also has the advantage that it is meaningfully different from other mathematical concepts, that is, non-ergodicity, that have separately appeared in the economics literature. This suggests that theories of path dependence should focus on how systems can exhibit long passage times out of local basins of attraction rather than multiple equilibria or multiple steady states per se. This in turn would suggest that analyses of path dependence should focus on understanding aggregate nonlinearities rather than the persistence of shocks as has occurred historically. The reason for this is that the second definition does not imply that actual shocks have persistent effects, only that they could have them.

The definitional ambiguities associated with path dependence are mirrored in the substantive discussions that have been developed concerning the economic environments in which it is supposed to occur. Liebowitz and Margolis (1990, 1995) have challenged David's claims about path dependence and the QWERTY keyboard and indeed have questioned the general empirical relevance of the concept. The Liebowitz and Margolis arguments, in the context of QWERTY, largely amount to claiming that there is no good evidence that the Dvorak keyboard is superior and that, further, the historical record indicates that the emergence of the QWERTY standard was driven by competitive forces to a much greater extent than acknowledged by David (rebuttals to their claims include David 2001). This debate has not been resolved and has generally been unproductive. As discussed in Durlauf (2005), the main problem is the lack of careful attention to microeconomic behaviours when analysing the historical evidence. For example, to the extent that evidence of path dependence is equated with the possible stability of a technologically inferior standard, then what matters in evaluating the claim is the level of information that was available to the individual economic actors when they made their standard choices, not what was *ex post* true. This requires much more explicit attention to the decision problems of the individual actors whose choices are collectively said to produce path dependence as well as the way in which an equilibrium configuration of choices occurs at each point in time.

Put differently, the path dependence literature has generally reasoned from aggregate observations towards microeconomic conclusions, whereas a rigorous formulation and empirical evaluation of path dependence as the property of an economic system requires that one start with individual decisions and reason towards aggregate implications. What this means is that resolution of whether network externalities, for example, have produced multiple steady states in a particular market should be understood as claims about the nature of individual decisions and how aggregate equilibria emerge from them. It is well known, as a theoretical matter, that broad claims such as the assertion that markets select for rationality, efficiency and the like (cf. Blume and Easley 1992) depend on details of the economic environment under study. By implication, formal microeconometric analysis will be necessary for empirical adjudication of claims concerning path dependence.

The existing microeconometric literature makes it very clear that there exist deep difficulties in determining whether path dependence is present in a given environment. For example, without strong assumptions on the decision rules of individual agents, one cannot identify whether the equilibrium of a given model is or is not unique. Indeed, even with individual level data, so that the decision rules of the agents can in principle be estimated, identification of whether an environment can produce multiple equilibria is difficult (cf. Brock and Durlauf 2008; Tamer 2003). From the econometric perspective, one basic problem is that any argument that an equilibrium has emerged that is not unique implicitly requires identification of the strength of the interdependences in individual choices, for example network effects. These interdependences cannot be identified unless one is willing to make assumptions about the correlated (across individuals) unobserved components to the costs and payoffs of the choices that are made; some possibilities on how to do this appear in Brock and Durlauf (2008). Further, even when there are no such correlated components, there are cases where the degree of interdependence cannot be identified when there are correlated observables components; this was established in Manski (1993) who calls this the reflection problem. At the current writing, *none* of these issues has been explicitly considered in the study of path dependence.

Thus, the current path dependence literature has had mixed success. From the perspective of the identification of interesting facts and the description of candidate environments for multiple steady states, the path dependence literature has been quite stimulating. From the perspective of developing a new theoretical view of economic outcomes, something that the more grandiose writings on path dependence sometimes allege they do, the literature is still highly imprecise and speculative. Thus the contributions of path dependence research really amount to the delineation of interesting historical episodes, episodes whose interpretation has yet to be resolved.

See Also

- ► Economy as a Complex System
- ► Path Dependence and Occupations
- ► Path Dependence in Technical Standards
- ► Social Interactions (Empirics)
- ► Social Interactions (Theory)

Bibliography

Arthur, W. 1989. Increasing returns, competing technologies and lock-in by historical small events: The dynamics of allocation under increasing returns to scale. *Economic Journal* 99: 116–131.

Blume, L., and D. Easley. 1992. Evolution and market behavior. *Journal of Economic Theory* 58: 9–40.

Blume, L., and S. Durlauf. 2001. The interactions-based approach to socioeconomic behavior. In *Social dynamics*, ed. S. Durlauf and H. Young. Cambridge, MA: MIT Press.

Brock, W., and S. Durlauf. 2008. Identification of binary choice models with social interactions. *Journal of Econometrics*.

Cowan, R. 1990. Nuclear power reactors: A study in technological lock-in. *Journal of Economic History* 50: 541–567.

David, P. 1985. Clio and the economics of QWERTY. *American Economic Review* 75: 332–337.

David, P. 1986. Understanding the economics of QWERTY: The necessity of history. In *Economic history and the modern economist*, ed. W. Parker. Oxford: Blackwell.

David, P. 2001. Path dependence, its critics and the quest for 'historical economics'. In *Evolution and path dependence in economic ideas: Past and present*, ed. P. Garrouste and S. Ioannides. Cheltenham/Northampton: Edward Elgar.

Durlauf, S. 2005. Complexity and empirical economics. *Economic Journal* 112: 225–243.

Farrell, J., and G. Saloner. 1985. Standardization, compatibility, and innovation. *RAND Journal of Economics* 16: 70–83.

Liebowitz, S., and S. Margolis. 1990. The fable of the keys. *Journal of Law and Economics* 22: 1–26.

Liebowitz, S., and S. Margolis. 1995. Path dependence, lock-in, and history. *Journal of Law, Economics, and Organization* 11: 205–226.

Manski, C. 1993. Identification of endogenous social effects: The reflection problem. *Review of Economic Studies* 60: 531–542.

Puffert, D. 2002. Path dependence in spatial networks: The standardization of the railway track gauge. *Journal of Economic History* 39: 282–314.

Tamer, E. 2003. Incomplete simultaneous discrete response model with multiple equilibria. *Review of Economic Studies* 70: 147–165.

Path Dependence and Occupations

Maristella Botticini and Zvi Eckstein

Abstract

Path dependence in occupations refers to the observed occupational distribution in a population or in a sub-population at a point in time that depends on changes that occurred years or centuries earlier. Path dependence in occupations can be the outcome of the cumulative concentration of certain productive activities in specific regions over time, it can emerge through the effect of parental income or wealth on offspring's occupations and incomes, or it can be the outcome of group effects. Some historical cases are selected to illustrate the various mechanisms through which path dependence in occupations can emerge or disappear.

Keywords

Affirmative action; Black–white labour market inequality (USA); Civil Rights Act 1964 (USA); Commercial and trade diasporas; Economic geography; Ghettoes; Great Migration (USA); High school movement (USA); Inequality (explanations); Innovation; Intergenerational income mobility; Intergenerational occupational mobility; Intergenerational skill transfers; International migration; Jewish economic history; Kuznets, S.; Labour market discrimination;

Manufacturing belt (USA); Migration and labour markets; Occupational choice; Occupational structure; Path dependence; Path dependence in occupations; Poverty traps; QWERTY keyboard; Religion and economic development; Residential mobility; Rural-urban migration; Urbanization; Women's work

JEL Classifications

J24

Path dependence in occupations can be interpreted to mean that the observed occupational distribution in a population or in a sub-population at a point in time depends on changes that occurred years or centuries earlier (for example, a war that siphons off certain types of workers, the enactment of anti-discriminatory labour practices, a technological invention which is not gender- or race-neutral). This definition is consistent with the notion of path dependence suggested in the economic history literature by David (1985) with the example of the 'standard QWERTY' keyboard. Under this definition one can include both the cases in which particular innovations in the economy have permanent consequences and those instances in which particular shocks are not self-correcting, so that they remain permanent in the absence of some countervailing change.

To show that there is path dependence in occupations, one has to describe the exact sequencing of events related to the initial change and show that they had a permanent effect on the occupational choice and distribution observed later. In other words, one has to show that, at a given point in time, multiple occupational distributions were available for selection, and theory is unable to predict or explain the occupational structure that will be chosen. Then, a change occurs and an occupational distribution is favoured over competing ones. Finally, the selected occupational structure capitalizes on initial advantage and is stably reproduced over time.

The economics literature identifies a number of possible sources of path dependence (see for discussions Arthur 1989; David 1994; Liebowitz and Margolis 1995; Blume and Durlauf 2005). For example, the economic geography literature explains path dependence in occupations as the outcome of the cumulative concentration of certain productive activities in specific regions over time (for example, Krugman 1991b; Fujita et al. 1999). This literature highlights the potentially big impact of increasing returns and cumulative processes, which in turn can make the role of historical accidents decisive. Small changes in the parameters of the economy may have large effects. For example, if transportation costs, economies of scale, and the share of non-agricultural goods in expenditure cross a critical threshold, population may start to concentrate and regions to diverge; once started this process will feed on itself.

However, increasing returns are not necessary for path dependent processes (Bowles and Gintis 2002). For example, in models of intergenerational mobility where individual-level characteristics matter (for example, Becker and Tomes 1979; Loury 1981; Banerjee and Newman 1991, 1993; Galor and Zeira 1993; Eckstein and Zilcha 1994; Mookherjee and Ray 2002, 2003), the existence or absence of path dependence in relative economic status across generations emerges through the effect of parental income or wealth on offspring's occupations and incomes.

In contrast, starting from the seminal work of Shelling (1971), in membership models an individual's economic choices are influenced not only by his or her traits but also by characteristics of the group of individuals with whom the person typically interacts (see Durlauf 2006, for a discussion of these models and related empirical literature). Groups may differ in average level of schooling, cognitive functioning, occupational structure and wealth level. Some groups are exogenously determined, for example by ethnicity or gender. Other groups are endogenously determined. For example, individuals may be strongly influenced by groups such as residential neighbourhood, the schools attended, and the co-workers at various jobs. Group effects on economic success are well documented and may arise for a number of reasons, including discrimination, conformist effects on behaviour, differential access to information, and complementarities in production.

An exhaustive survey of historical and contemporary examples of path dependence in occupations is beyond the scope of this article. Instead, we selected some examples which illustrate the various mechanisms discussed above through which path dependence in occupations can emerge or disappear.

Jewish Economic History in the Past Two Millennia

At the beginning of the first millennium, an exogenous change in the religious and social norm that defined Judaism occurred as a result of the shift in the leadership within the Jewish community. Before the destruction of the Temple in Jerusalem in 70 CE, the Jewish population in Eretz Israel, which consisted mainly of farmers, was segmented in many religious groups. After the destruction of the Temple, many Jewish sects disappeared, whereas the Pharisees became the dominant group. They replaced sacrifices with the study of the Torah in the synagogue. The transformation of the religion created the need for the devoted Jews to be literate and to educate their children. In about 200 CE, the transformation of Judaism reached its full-fledged stage with the compilation of the Mishna. Also, a new social norm came to prevail according to which an illiterate Jewish individual was considered an outcast in the community.

Despite education being very costly and ‘useless’ in production for farmers, religious instruction and primary education became more and more widespread among the Jewish communities in Eretz Israel and Babylonia from the second to the seventh century. The spread of literacy among the Jewish rural population is even more impressive when compared with the literacy rates of the non-Jewish rural population in the same period. In the Roman, Byzantine, Christian and Persian worlds there was no mandatory primary education, and the non-Jewish rural population was almost entirely illiterate.

Before 400 CE almost all Jews in the three main centres of Jewish life in the classical period – Palestine, Babylonia, and Egypt – were farmers, exactly like the rest of the population. The transition away from agriculture into crafts, trade, and moneylending started in the Talmudic period (200–500 CE), especially in Babylon. In the fifth and sixth centuries, some literate Jews abandoning agriculture moved to the towns and became small shopkeepers, craftsmen and artisans. However, given the stagnant economies in the late Roman, early Byzantine and Persian empires in the fourth to the seventh centuries, the growing number of literate Jewish farmers could not find skilled occupations in the existing cities at that time and many of them converted out of Judaism. World Jewry was reduced from about 4.5 million in 70 CE to about 1.5 million in 600 CE, with 80 per cent living in Babylonia.

But in the eighth and ninth centuries, another exogenous event occurred: massive urbanization in the newly established Muslim empire under the Abbasid caliphate vastly increased the demand for urban, skilled occupations. The *literate* Jewish rural population in Iraq and later in the Abbasid empire as a whole moved to urban centres, abandoned agriculture, and became engaged in a wide range of crafts, local and long-distance trade, moneylending, tax-farming and the medical profession. This occupational transition took about 150 years, and by 900 CE almost all Jews in Iraq, Persia, Syria and Egypt, were engaged in urban occupations. In contrast, most non-Jews remained farmers, even though they could engage in any occupation in the regions under Muslim rule. These two facts identify the educational reform in Judaism around 200 CE as the key factor for the occupational transition of the Jewish people (Botticini and Eckstein 2005).

Judaism, with its costly religious norm regarding education, can thrive in the long run only if the Jews can find occupations in which their earnings significantly gain from literacy (Botticini and Eckstein 2006). The voluntary diaspora of the Jews to western Europe during the tenth to the 13th centuries, to eastern Europe in the 16th and 17th centuries, and then worldwide supports this argument. Other minorities within the Muslim empire under the Abbasid caliphate did not migrate to western Europe even though no prohibitions prevented them from doing so. The distinctive engine of the Jewish migrations to the West was the incentive to maximize the returns to their investment in education. Hence, these two facts identify the link between the 'historical accident' and the voluntary diaspora of the Jews in search of urban, high-skill and high-income occupations.

The large Jewish population of Iraq and Iran, which amounted to about 800,000 in 1250, almost disappeared when the Mongol invasions brought the Near East back to a subsistence farming economy. In contrast, the small Jewish population in Europe survived, kept its literacy and educational distinctiveness, and through urban and skilled occupations reached high standards of living.

These urban, skilled occupations remained the distinctive mark of the Jewish people throughout their history, as clearly highlighted by the data provided by Kuznets (1960): in the countries which hosted the largest Jewish communities in the early 20th century (countries in eastern Europe, Russia, the United States and Canada), between 96 and 99 per cent of the Jews were engaged in non-agricultural occupations even though no restrictions prevented them from being farmers. Chiswick (2005) documents the same occupational selection of the Jewish population in the United States as late as the year 2000. For example, about 53 per cent of adult Jewish men are engaged in professions such as law, medicine, and academia, whereas the percentage for white non-Jewish men is about 20 per cent. In contrast, only six per cent of adult Jewish men are employed in the construction, transportation, and production sectors in comparison with about 39 per cent of adult non-Jewish men.

Jewish economic history fits very well the multiple features of path dependence outlined in the introduction. On the one hand, two exogenous changes (the transformation of the religious norm in the first and second centuries CE and the urbanization in the Muslim empire in the eighth and ninth centuries) created a permanent effect on the occupational distribution among the Jews. On the other hand, the mechanisms through which these changes worked to affect the occupational structure of the Jews in the long run were twofold: the intergenerational transmission of skills and literacy from parents to children, and the peer pressure (social penalty) that the Jewish communities imposed on those who did not invest in their children's education.

Commercial and Trade Diasporas

As membership models would predict, ethnic groups can influence the occupational distribution of immigrants in a country and create occupational clustering by ethnicity. One of the most visible examples of this occupational clustering is offered by the so-called commercial and trade diasporas.

A diaspora is any ethnic group without a territorial base within a given polity, and whose social, economic and political networks cross the borders of nation states. In particular, trade and commercial diasporas are those diasporas whose members specialize in trade and commercial activities or, more generally, in urban, skilled jobs. Historical examples include the Jews in the last two millennia, the Parsi (Zoroastrian) diaspora from Iran, the Huguenots in early modern and modern western Europe, the Armenians, the Greeks of the Ottoman Empire, the Germans throughout eastern Europe in modern times, the Chinese in many areas of south-east Asia from the 15th to the 20th century, the Indian middleman minorities of east Africa and Malaya, the Pakistanis in Great Britain, and the Lebanese Christians in 18th-century Egypt and contemporary west Africa (Botticini 2003).

Commercial and trade diasporas – indeed, diasporas in general – have been characterized by strong linguistic skills, often including the ability to speak and write in both their own and alien languages. This enabled members of a diaspora to maintain communication networks within the group and to use alien languages for practical purposes. Maintaining the common original language is one of the means to enhance the organization of a diaspora. Others mechanisms include the establishment of communal institutions, such as the commercial coalitions among the Jews in the Mediterranean in the high Middle Ages (Greif 1989) or the Chinese societies known as *Houei;* the development of a common set of commercial laws or norms whose enforcement is delegated to courts within the communities; and strong endogamous marriage strategies.

In some cases, exogenous changes have created or reinforced occupational selection among ethnic or religious groups. For example, it has been often argued that legal prohibitions and the exclusion of Jews from guild membership in medieval and early modern Europe would account for their occupational selection into moneylending and the medical profession. Similarly, it has been pointed out that, after the revocation of the Edict of Nantes by King Louis XIV in 1685 that made Protestantism illegal, many Huguenots (French Protestants) emigrated to Ireland, England, Prussia, and America, where they contributed to the development of industries and trades. The Agricultural Law of 1870 in Indonesia against land ownership by ethnic Chinese has been cited to explain the exclusion of the Chinese diaspora from farming and agricultural activities.

In other instances, the occupational distribution was altered by rulers who substituted one diaspora for another if they perceived the change to be advantageous for them. Thus, in the Ottoman Empire, Catholic Levantines, who held the leadership in crafts and trade in the 15th century, were replaced by the Jews in the 16th and 17th centuries, followed by the Greeks until the beginning of the 19th century and Armenians during the 19th century.

Geography also played a role in the occupational specialization of some ethnic groups. With the European geographical expansions and the establishment of colonial rule in south-east Asia and west and east Africa during the 19th and 20th centuries, Lebanese Christians, Chinese, and Indians have contributed to the establishment of commercial economies in the European colonial empires.

The Manufacturing Belt in the United States

The establishment and remarkable persistence of the manufacturing belt in the United States is one of the most prominent examples of geographic concentration which in turn affected the occupational distribution of the US population.

Early in the history of the United States, when most of the population was engaged in agriculture, when transportation costs were high, and when manufacturing was characterized by few economies of scale, no concentration could occur. When the United States started to industrialize, manufacturing first developed in regions where most of the agricultural population outside the South was located. The manufacturing belt developed in the second half of the 19th century when economies of scale in manufacturing increased, transportation costs fell, and the share of the population in non-agricultural occupations rose. The initial advantage of the manufacturing belt was locked in, leading the bulk of US manufacturing to be concentrated in a relatively small part of the north-east and the eastern part of the Midwest. It persisted even as the centre of gravity of agricultural and mineral production shifted to the West. As late as 1957, the manufacturing belt still contained 64 per cent of US manufacturing employment (Krugman 1991a).

Intergenerational Occupational Mobility in Britain and the United States Since 1850

Unlike today, the United States in the 19th century was ‘exceptional’ in the occupational mobility experienced by its population (as well as in its geographic mobility) compared with Europe. As documented by Long and Ferrie (2005), this contrast is even more striking when 19th-century United States is compared with 19th-century Britain – the country with which it shared legal traditions and property rights systems and sources of labour, capital, and technology.

Differences have been attributed to a number of factors. First, the absence of feudalism and of strong craft guilds has been put forth as one reason for the higher occupational mobility in the United States. Second, at least some of the high mobility in 19th-century United States may result from it being at an earlier stage of development than 19th-century Britain, so its farm sector was relatively larger.

Third, the United States provided considerably more public education than Britain in the middle of the 19th century: the primary school enrolment rate was one and a half times greater in the United States than in Britain. The US educational system in the second half of the 19th century, though less extensive at the secondary and post-secondary levels than European systems, was considerably more egalitarian (Goldin and Katz 2003). To the extent that intergenerational mobility is greater where fewer parents are wealth-constrained, superior mobility in the United States may well have been a consequence of its educational system, which provided a public alternative to a private education that was outside the reach of many families.

Fourth, residential mobility to places that were growing more rapidly than others may have provided an alternative to direct investment in human capital. Cities (such as Chicago) sprang up initially to provide services demanded as the frontier expanded. Though US labour markets in the North were well-integrated at the regional level by the middle of the 19th century, differences across smaller units of geography may have continued to present opportunities for 'locational arbitrage' that provided a route to occupational change through the start of the 20th century (Long and Ferrie 2005).

The Feminization of Teaching and Clerical Work in the United States

Teaching Profession

Today in the United States the vast majority of elementary and secondary teachers are women. In 2000, the female proportion among teachers was 76 per cent. Much earlier in American history, however, this was not the case. The feminization of teaching occurred over the course of the 19th century and continued throughout the 20th century. Two exogenous factors changed the social norm and attitude towards female teachers in the United States and, therefore, significantly contributed to the feminization of teaching: (*a*) the ethnic, national, and cultural identity of the European settlers who established their communities in the Northern, Midwestern, and Southern states, and (*b*) the wars (especially the American Civil War and the First World War).

Relatively early in the 19th century, women came to dominate teaching in New England through the establishment of two educational institutions: the so-called 'dame schools' and a two-tier system divided into winter and summer sessions. The 'dame schools' were an educational institution imported by British settlers, in which women taught very young children as they were considered the natural carers for these children. The division into winter and summer sessions reinforced this gender-specific assignment of teachers to pupils according to age. As winter sessions were geared towards older boys, male teachers were considered to have greater human capital and skills to enforce discipline among them. Female teachers were considered better equipped to teach summer sessions attended by younger children. As population spread westward in the North, the female percentage in teaching increased in these states (Carter and Margo 2007).

In contrast, because of the different ethnic and national background of the European settlers who established themselves in the US South, neither 'dame schools' nor the two-tier system were developed and the percentage of female teachers remained much lower there until the Civil War. But even within the North itself, the role of culture and institutions in affecting the gender distribution in the teaching profession is illustrated by regional variation. In Illinois counties where the settlers were mainly Yankees, female teachers were quite common, whereas in those counties where the settlers were mainly Southerners, male teachers predominated (Carter and Margo 2007).

As Perlmann and Margo (2001) have shown, the American Civil War significantly contributed to the feminization of teaching. In 13 Northern and Midwestern states, the average share of female teachers rose from about 57 per cent in 1860 to 67 per cent in 1865 and 79 per cent in 1915. During the war women took jobs in teaching, substituting for men who were at war. When the war ended, there was some mean reversion, but not back to the original equilibrium.

The entry of many women into teaching during the Civil War changed the social norm and attitudes toward female teachers by making the bias against them gradually fade. In the earlier decades, the argument against hiring female teachers had been that, especially in winter classes when adult boys attended school, women lacked the skills to discipline these students. However, the entry of women in the teaching profession during the war to substitute for the male teachers gave them the opportunity to show that they could be as effective as their male colleagues in teaching and maintaining the discipline among students. This changed the social norm and attitude toward hiring female teachers, which increased the feminization of teaching in both the Northern states and in the South, where the share of female teachers reached unprecedented levels, rising from about 35 per cent in 1875 to 73 per cent in 1915 (Perlmann and Margo 2001, p. 169).

The First War World had a similar effect on the selection of women into the teaching profession, although on a smaller scale. After the Second World War, women entered many other occupations and professions. Yet the predominance of female teachers in primary and secondary schools holds to the present day.

Clerical Work

In 1870, fewer than three per cent of all clerical employees were women. In 1930, women made up over half (52.5 per cent) of the total clerical workforce, and today the clerical sector is one of the major employers of women. The most rapid increases occurred in two decades, 1880–90 and 1910–20, as the outcome of two exogenous shocks on the demand side of the labour market coupled with a profound transformation on the supply side of the same market.

On the demand side, Rotella (1981) has argued that the adoption and diffusion of the typewriter in the 1880s, the growth of large firms and the expansion of the government sector in the 1910s created a huge demand for clerical work. Specifically, the diffusion of the typewriter made the skills required of clerical labour no longer firm-specific, as it had been when employers preferred to hire male workers who were expected to have a long working life within the firm. With the development of the modern, mechanized office, employers could afford to hire young, educated women who had high expected turnover and who desired clean, high-status employment. Later, in the 1910s, the growth of large firms and the expansion of the government sector through regulation and tax laws greatly increased the demand for information and information processing within firms and government offices. Again, this shift in demand was not gender neutral: it favoured women, and women came to dominate office work, basically after about 1910.

On the supply side, Goldin (1986, 1990) has shown that the huge increase in high school attendance around the turn of the 20th century – the so-called ‘high school movement’ – dramatically increased the supply of young, educated women in the labour market. These women offered a relatively cheaper and easier to monitor labour force.

The Occupational Transition of African-Americans in the 20th Century

After the American Civil War, a steady stream of African-Americans moved out of the South to the North. It has been estimated that from 1870 to 1910 about 535,000 blacks emigrated from the South on the net as the outcome of the large wage differentials between the North and the South and of the increased human capital acquired by the first generations of blacks after the Emancipation (Margo 1990, ch. 7). This migration, though, did not have a huge impact on the overall occupational and residential distribution of African-Americans. In fact, in 1900 approximately 90 per cent of the blacks still lived in the South, and the majority of them worked in agriculture and were very poor.

In contrast, from 1910 to 1950 the Great Migration brought about 3.5 million African-American people out of the South mainly to the urban North. Even when migration occurred to rural areas in the North, it invariably involved a shift out of agriculture. The Great Migration represented a watershed in African-American economic history and implied a profound and permanent transformation of the occupational distribution of the blacks in the United States.

The relevant exogenous shocks that fuelled both the Great Migration and the permanent change in the occupational distribution of the blacks were the two world wars and a combination of government policies.

First, quotas set by the US government on foreign immigration after the First World War greatly accelerated the outmigration of blacks from the South in the 1920s, as blacks were substitutes for the foreign-born immigrants (Collins 1997).

Second, the Second World War was an even bigger exogenous shock. When the United States entered the war, demand for white workers in the war-industry sector increased at the same time as the military was siphoning off potential workers. US employers were faced with a tough choice: either to follow the prevailing taste for discrimination among employers and white workers and the social norm against hiring black workers, or to expand production and to gain profits by hiring black workers.

The enforcement by President Roosevelt of the anti-discriminatory policy amongst defence contractors through the Fair Employment Practice Committee (FEPC) established in 1941 was the exogenous change in government policy that helped employers choose the second option and hire black workers despite the prevailing taste for discrimination (Collins 2001). The impact of this government intervention was twofold. It made defence contractors hire black male workers who otherwise would not have been hired because of the hostility of white male workers towards hiring fellow black workers. At the same time, it started changing the social norm and attitude against hiring black workers in other firms and industries in those instances when the enforcement of the anti-discriminatory policy among the defence contractors sector had spillover effects on other firms' hiring practices.

The combination of the two exogenous shocks – the Second World War and the establishment of the Fair Employment Practice Committee – had a large impact on the occupational and residential transition of African-Americans. Between 1940 and 1950 the proportion of black male workers classified as operatives (semi-skilled) rose from 12.6 to 21.4 per cent, and the proportion in manufacturing industries rose from 16.2 to 23.9 per cent (Collins 2000). This transition into manufacturing and war-related industries greatly contributed to the economic progress of blacks, as the data on the substantial wage premium these workers earned indicate.

A similar effect occurred ten years later as the outcome of another major change in government policy. The *Brown vs Board of Education* Supreme Court's decision in 1954, which invalidated school segregation in the US South, the enactment of Title VII of the 1964 Civil Rights Act, which forbade discrimination in employment, the establishment of the Office of Federal Contract Compliance (OFCC), which monitored the anti-discrimination and affirmative action responsibilities of government contractors, and the passage of the Voting Rights Act of 1965 were the most famous among several government policies designed to eliminate discrimination against blacks. Donohue and Heckman (1991) show that a significant portion of the sustained improvement in the labour market status of black males from 1965 to 1975 (especially in the US South) was the outcome of these changes in government policies.

Poverty Traps

Intertemporal social interactions (that is, social interactions in which choices made at one time affect others made later) can create path dependence in occupations through a variety of mechanisms. Role models and peer group effect models are two examples of these mechanisms. Suppose, as role models do, that the decision to attend college by a young adult depends on the percentage of college graduates among adults in his community. Then two communities, one where the adults are all college graduates and the other where none are, can converge to different levels of college attendance in a steady state, leading to path dependence in occupational and economic segregation across long time periods and generations.

The persistence of ghettoes and poverty traps are the two most visible examples of the intertemporal effect of group membership on individual outcomes (Bowles et al. 2006). Poverty traps are situations where the evolution of individual wealth is governed by a path-dependent process such that, depending on initial conditions, otherwise identical individuals or groups (ethnic, linguistic, religious) may remain for long periods of time ‘locked into’ poverty. The key characteristic of a poverty trap is that the ‘good’ and ‘bad’ outcomes are self-enforcing, so that small interventions or chance events will not alter the long-term outcome. Recent evidence of the persistence of income differences between races, even after some of the structural determinants of inequality (such as colonialism, inequalities of educational opportunity, and de jure segregation) have been removed, point to the importance of historical contingency and ‘lock-in effects’ in the process that generates inequality (Loury 2002; Bowles 2006).

See Also

- ▶ Ghettoes
- ▶ Income Mobility
- ▶ Inequality Between Nations
- ▶ Inequality (Global)
- ▶ Inequality (International Evidence)
- ▶ Inequality (Measurement)
- ▶ Intergenerational Income Mobility
- ▶ Path Dependence
- ▶ Social Networks in Labour Markets
- ▶ Social Norms

Bibliography

Arthur, W. 1989. Increasing returns, competing technologies and lock-in by historical small events: The dynamics of allocation under increasing returns to scale. *Economic Journal* 99: 116–131.

Banerjee, A., and A. Newman. 1991. Risk-bearing and the theory of income distribution. *Review of Economic Studies* 58: 211–235.

Banerjee, A., and A. Newman. 1993. Occupational choice and the process of development. *Journal of Political Economy* 101: 274–298.

Becker, G., and N. Tomes. 1979. An equilibrium theory of the distribution of income and intergenerational mobility. *Journal of Political Economy* 87: 1153–1189.

Blume, L., and S. Durlauf, ed. 2005. *The economy as an evolving complex system III: Current perspectives and future directions*. Oxford/New York: Oxford University Press.

Botticini, M. 2003. Commercial and trade diasporas. In *Oxford encyclopedia of economic history*. New York: Oxford University Press.

Botticini, M., and Z. Eckstein. 2005. Jewish occupational selection: education, restrictions, or minorities? *Journal of Economic History* 65: 922–948.

Botticini, M., and Z. Eckstein. 2006. *From farmers to merchants, voluntary conversions and diaspora: A human capital interpretation of Jewish history*, Discussion Paper No. 5571. London: CEPR.

Bowles, S. 2006. Institutional poverty traps. In *Poverty traps*, ed. S. Bowles, S. Durlauf, and K. Hoff. Princeton: Princeton University Press.

Bowles, S., S. Durlauf, and K. Hoff, ed. 2006. *Poverty traps*. Princeton: Princeton University Press.

Bowles, S., and H. Gintis. 2002. The inheritance of inequality. *Journal of Economic Perspectives* 16(3): 3–30.

Carter, L., and R. Margo. 2007. Feminization of teaching in the United States. In *Gender and education in the United States*, ed. B. Bank. Westport: Greenwood Press.

Chiswick, R. 2005. *The occupational attainment of American Jewry: 1990–2000*, IZA Discussion Paper No. 1736. Bonn: Institute for the Study of Labour.

Collins, W. 1997. When the tide turned: immigration and the delay of the Great Black migration. *Journal of Economic History* 57: 607–632.

Collins, W. 2000. African-American economic mobility in the 1940s: A portrait from the Palmer Survey. *Journal of Economic History* 60: 756–781.

Collins, W. 2001. Race, Roosevelt, and wartime production: Fair employment in World War II labor markets. *American Economic Review* 91: 272–286.

David, P. 1985. Clio and the economics of QWERTY. *American Economic Review Papers and Proceedings* 75(2): 332–337.

David, P. 1994. Why are institutions the ‘carriers of history’? Path dependence and the evolution of conventions, organizations, and institutions. *Structural Change and Economic Dynamics* 5: 205–220.

Donohue, J. III, and J. Heckman. 1991. Continuous versus episodic change: the impact of the civil rights policy on the economic status of blacks. *Journal of Economic Literature* 29: 1603–1643.

Durlauf, S. 2006. Groups, social influences and inequality: a membership theory perspective on poverty traps. In

Poverty traps, ed. S. Bowles, S. Durlauf, and K. Hoff. Princeton: Princeton University Press.

Eckstein, Z., and I. Zilcha. 1994. The effects of compulsory schooling on growth, income distribution and welfare. *Journal of Public Economics* 54: 339–359.

Fujita, M., P. Krugman, and A. Venables. 1999. *The spatial economy: Cities, regions, and international trade*. Cambridge, MA: MIT Press.

Galor, O., and J. Zeira. 1993. Income distribution and macroeconomics. *Review of Economic Studies* 60: 35–52.

Goldin, C. 1986. Monitoring costs and occupational segregation by sex: A historical analysis. *Journal of Labor Economics* 4: 1–27.

Goldin, C. 1990. *Understanding the gender wage gap: An economic history of American women*. Chicago: University of Chicago Press.

Goldin, C., and L. Katz. 2003. *The 'virtues' of the past: Education in the first hundred years of the new republic*, Working Paper No. 9958. Cambridge, MA: NBER.

Greif, A. 1989. Reputation and coalitions in medieval trade: Evidence on the Maghribi traders. *Journal of Economic History* 49: 857–882.

Krugman, P. 1991a. History and industry location: The case of the manufacturing belt. *American Economic Review Papers and Proceedings* 81(2): 80–83.

Krugman, P. 1991b. Increasing returns and economic geography. *Journal of Political Economy* 99: 483–499.

Kuznets, S. 1960. Economic structure and life of the Jews. In *The Jews: their history, culture, and religion*, ed. L. Finkelstein. Philadelphia: Jewish Publication Society of America.

Liebowitz, S., and S. Margolis. 1995. Path dependence, lock-in, and history. *Journal of Law, Economics, & Organization* 11: 205–226.

Long, J., and J. Ferrie. 2005. *A tale of two labor markets: Intergenerational occupational mobility in Britain and the United States since 1850*, Working Paper No. 11253. Cambridge, MA: NBER.

Loury, G. 1981. Intergenerational transfers and the distribution of earnings. *Econometrica* 49: 843–867.

Loury, G. 2002. *The Anatomy of Racial Inequality*. Cambridge, MA: Harvard University Press.

Margo, R. 1990. *Race and schooling in the South, 1880–1950: An economic history*. Chicago: University of Chicago Press.

Mookherjee, D., and D. Ray. 2003. Persistent inequality. *Review of Economic Studies* 70: 369–393.

Mookherjee, D., and D. Ray. 2002. Contractual structure and wealth accumulation. *American Economic Review* 92: 818–849.

Perlmann, J., and R. Margo. 2001. *Women's work? American schoolteachers, 1650–1920*. Chicago: University of Chicago Press.

Rotella, E. 1981. *From home to office: US women at work, 1870–1930*. Ann Arbor: UMI Research Press.

Shelling, T. 1971. Dynamic models of segregation. *Journal of Mathematical Sociology* 1: 143–186.

Path Dependence in Technical Standards

Douglas Puffert

Abstract

The value of standardization leads to the persistence of established technical practices despite, in some cases, imperfect adaptation to current exogenous conditions. Economic explanation of these practices therefore requires reference to history. The conditions leading to path dependence in technical standards, as well as path independence, are examined with reference both to economic theory and to case studies of the QWERTY keyboard, videocassette recorder systems, railway track gauge and other railway standards. The controversy over path dependence is discussed.

Keywords

Increasing returns to adoption; Information technology; Learning by doing; Learning by using; Learning effects; Multiple equilibria; Network effects; Path dependence; Path dependence in technical standards; QWERTY keyboard; Railway track gauge; Standards; Technical interrelatedness; Technical standards; Transaction costs

JEL Classification

O33; N70; C63

Path dependence is the dependence of outcomes on the course of previous outcomes, and thus on past conditions, rather than simply on current exogenous conditions. In a path-dependent process of economic change, choices motivated by transitory conditions can have results that persist long after those conditions change. The early conditions that have persisting effects could be systematic in nature, but the literature has focused more on the role of non-systematic 'small' events in selecting one potential path of later outcomes rather than another. A path-dependent process is non-ergodic, that is, its limiting distribution of possible outcomes changes as a function of its specific, evolving history. History matters for later outcomes, and economic explanation is incomplete without accounting for that history.

Early choices can have particularly strong effects in the context of technologies that exhibit ‘increasing returns to adoption’ (Arthur 1989), in that specific practices become more valuable to each user as the total number of users rises. These increasing returns often give rise to standards – rules or practices that enable adopters to pursue some sort of value-producing interaction with other adopters. For example, railway companies that adopt a common track gauge (width between the rails) can exchange cars without reloading, while typists who learn a standard keyboard layout can apply their skills in any office that uses standard machines. Increasing returns to adoption can arise either on the demand side or the supply side of a market. On the demand side, as in the cases of railway gauges and typewriter keyboards, adopters gain from participation in physical or virtual networks of adopters. On the supply side, learning effects – learning by doing or learning by using – reduce the cost or improve the characteristics of a product as cumulative adoption increases.

Quite often, a technology embodying increasing returns to adoption offers a range of specific practices that could form the basis for a standard. Railways have used track gauges ranging from about 600 mm (two feet) to 2140 mm (seven feet), and typewriter or computer keyboards could use 26-factorial (about 1026) different orderings of the letters. These practices may represent different, diverging, potential paths of outcomes. Once early adopters choose a particular practice, later adopters have an incentive to match those choices in order to gain the benefits of compatibility. Thus, increasing returns can give rise to positive feedbacks among agents’ choices. The selection process (or allocation process) that results does not converge to a unique equilibrium outcome. Rather, such a process has multiple potential equilibria or, rather, equilibrium paths. These equilibria could vary substantially both in their general efficiency (total payoffs) and in their distribution of payoffs among different agents. Which equilibrium is selected depends in large part on early choices.

Requirements for Path Dependence

Two things are necessary for path dependence to make a difference for outcomes (David 1999, 2001). First, the conditions or criteria that determine early choices – and thus one branching path rather than another – must not be closely correlated with the conditions or interests that matter later. Second, the selected path of outcomes must constitute a locally stable equilibrium, so that the selection process does not simply revert to an outcome that is determined by later conditions or interests.

Empirically, the most important reasons that early choices might not reflect later conditions are limited information and limited technical capability. These are common occurrences in the early stages of a new technology. Innovators must often engage in exploratory behaviour to learn the possibilities for both technological development and market application (Nelson and Winter 1977), and the effective choice of a standard practice may precede much of this learning. For example, the standard track gauge used in most of the world today was chosen during the 1820s, when railway cars were little more than road wagons and when locomotives were little more than small steam engines set on wagons and linked by crank to a wheel. The gauge was not optimized for what railways were soon to become, let alone for what they are 180 years later. Many engineers since the 1830s have believed that a broader gauge would be more efficient for most purposes (Puffert 2002, 2008). As another example, the ‘QWERTY’ standard typewriter keyboard was developed by rearranging letter sequences so as to minimize the jamming propensities of one short-lived typewriter design in 1873. Modern eight-finger typing methods emerged a decade later (David 1986), and keyboards designed for such methods offer about a 10% improvement over QWERTY in typing efficiency (Norman 1990).

A further reason that early choices might not reflect later conditions and interests is that later adopters may have different interests with regard to the content of a standard than early adopters, but high transaction costs (or simple lack of foresight) may prevent the later adopters from influencing the early choices that determine their later options. In addition, the discounting of future payoffs would lead even perfectly foresighted agents to place little value on distant outcomes.

The second requirement for path dependence to have consequential effects, again, is that a selected path of outcomes be locally stable. Part of the cause of such stability is increasing returns to adoption. Increasing returns lead new adopters to adopt a practice simply because it is the established standard with a large installed base, even if they would prefer some alternative practice if that were to have a comparable installed base. In such an instance the established standard is said to be ‘locked in’, both in the economics literature (Arthur 1989) and in the business world.

Another part of the reason for the stability of a path of outcomes is switching costs – the hardware conversion costs, retraining costs, and transaction costs (that is, information and coordination costs) entailed in converting from an established standard practice to a superior alternative. Users are often restrained from converting not only by irreversible investments in the established practice but also by the technical interrelatedness of system components (David 1986), which makes piecemeal conversion impractical. In a railway, for example, individual equipment and fixtures of one gauge cannot simply be replaced with items of another gauge when they wear out. Rather, new equipment must continue to match the installed base of old equipment, and a conversion requires that all equipment be converted together. Furthermore, the value of compatibility may mean that any conversion must be coordinated among many agents.

These two requirements are by no means always present as technical standards are formed. Foresight into the technological and market opportunities of a new technology is often sufficient to enable early adopters or product sponsors to choose, in effect, a superior path of later outcomes. As an example, Sony and Philips introduced the standard compact disc (CD) format in the early 1980s after digital audio sampling theory, other relevant technologies, and market requirements were already well known. The standard has served quite well. In such instances path dependence, as such, plays no role in the selection process.

Furthermore, if switching costs are sufficiently low, then a less preferred path does not constitute a stable equilibrium. Thus, the potential inefficiency of a path-dependent process is generally limited to the cost (including transaction costs) of carrying out a remedy for this inefficiency. A less preferred path of outcomes may become unstable through innovations or market developments that either reduce the costs or increase the benefits of transition to a preferred practice (Puffert 2004). For example, invention of the low-cost rotary electrical converter in the 1880s helped bring an end to regional lock-in to DC electrical power by facilitating the coupling of AC transmission networks with applications that required DC (David 1991). Similarly, in contemporary information technology, adapters or ‘gateways’ arise frequently to link otherwise incompatible networks (David 1987). Sometimes these techniques offer a migration path from an inferior or obsolete practice to a superior one, making the selection process ‘path independent’.

The Controversy Over Path Dependence

The concept of path dependence first gained widespread attention in economics through Paul David’s (1985, 1986) interpretation of the case of QWERTY and through a series of theoretical discussions by W. Brian Arthur (1989, 1994). David’s thinking grew out of an earlier literature on how technical interrelatedness can inhibit adaptation to changing conditions (Veblen 1915; Frankel 1955; Kindleberger 1964; David 1975). Arthur combined mathematical models of non-ergodic processes in the natural sciences with economic theory concerning how increasing returns can give rise to multiple equilibria.

Arthur developed models in which stochastic fluctuations in the market shares of alternative products or practices are magnified by positive feedbacks until one practice gains the whole market, becoming locked in as a de facto standard. In view of the subsequent controversy over path dependence, it is worth noting that Arthur's (1989) primary model used two key assumptions that obviated the need to consider expectations and forward-looking behaviour. First, he assumed that alternative competing practices are unsponsored, rather than promoted by suppliers. Second, he assumed that increasing returns to adoption are based simply on learning effects embodied in a practice at the time of adoption, so that each adopter's payoffs depend only on the number of previous adoptions, not the number of future adoptions. Arthur discussed only briefly how outcomes of his model would differ under alternative assumptions. He acknowledged that, if increasing returns were based on network effects rather than learning effects, then each adopter's payoffs would continue to rise after adoption as the number of adopters increased. He reasoned that expectations would then lead to earlier lock-in, but he did not carry his analysis further.

Stan Liebowitz and Stephen Margolis (1990, 1995) raised a substantive critique of David's and Arthur's writings, based partly on exploring the implications of assumptions other than those of Arthur's model. The central thrust of their argument was that purposeful, profit-seeking, forward-looking behaviour can override the mechanisms of path dependence whenever, in their view, outcomes truly matter. According to Liebowitz and Margolis (1995), if agents can foresee that some potential future outcomes offer higher payoffs than others, then they have a variety of means to steer the selection process toward the preferred outcomes. Suppliers of products that embody superior practices can profit by promoting those practices to become standards. Adopters can also conduct transactions among themselves, by direct communication or market mechanisms, to assure that they realize the highest available payoffs. According to Liebowitz and Margolis, if means such as these are unable to realize a putatively superior outcome, then that is only because the costs (including transaction costs) of pursuing that outcome are greater than the benefits. In other words, they argued, the putatively superior outcome is not really superior. Agents may come to regret that earlier choices, made in the absence of good foresight, had made some conceivable outcome unattainable. However, Liebowitz and Margolis argued, such regret is naive, a crying over spilt milk.

Liebowitz and Margolis concluded that path dependence is likely to affect only features of the economy that no economic agent has a real reason to care about – and that are not worth much attention from economists or economic historians. They set forth a taxonomy of 'degrees' of path dependence: first degree, in which alternative outcomes have no consequences for efficiency; second degree, in which different outcomes offer differing payoffs but imperfect foresight and transaction costs prevent purposeful behaviour to attain the highest payoffs; and third degree, in which there is sufficient foresight and scope for forward-looking behaviour to attain the superior outcome, but this outcome is somehow still not attained. Liebowitz and Margolis argued that only the third type of path dependence would offer a real challenge to what they called 'the neoclassical model of relentlessly rational behavior leading to efficient, and therefore predictable, outcomes'. They claimed, however, that this type is unlikely to arise empirically.

David (1997, 1999, 2001) responded that Liebowitz and Margolis had mischaracterized several of the issues at stake. Puffert (2000, 2002, 2004, 2008) responded to the critics by incorporating the issues of foresight and forward-looking behaviour explicitly into models and case studies, and he argued that such behaviour is fully compatible with path dependence. He maintained that Liebowitz's and Margolis's taxonomy of 'degrees' is incomplete, leaving out a great range of cases where agents are neither fully passive nor fully able to control outcomes. Such cases demonstrate a rich, complex interplay between forward-looking behaviour and the legacy of past events.

Although David and Arthur had not sufficiently examined the issues of foresight and forward-looking behaviour, they also had not fully neglected these matters. David (1986), indeed, attributed path dependence in typewriter keyboards to the lack of perfect futures markets. Kenneth Arrow stated in his foreword to Arthur's collected articles that much of Arthur's analysis applies specifically where 'expectations are myopic, based on limited information' (Arthur 1994). This is not how Arthur himself explicitly interpreted his models, but it is how he applied them. For example, Arthur (1989), as well as David (1987), argued that path dependence may be particularly relevant for policy when early information is imperfect. Government, they argued, might improve information and later outcomes by exploring the potential payoffs of alternative practices before one practice became locked in. Such a policy later proved its value when the US government sponsored a competition among alternative high-definition television systems, resulting in the accelerated development of a superior digital technology while preventing lock-in to a soon-to-be outmoded analog system.

The relevance of all these considerations must be judged empirically. We begin with the disputed case of QWERTY.

The QWERTY Keyboard

David (1985, 1986) argued that QWERTY gained a lead over rival keyboard systems due to the happenstance that instruction in eight-finger 'touch' typing was developed first for QWERTY during the mid-1880s. The best-trained typists used QWERTY, so office managers hired them and bought QWERTY machines to match. This, in turn, gave budding typists, typing schools, the writers of typing manuals, and typewriter manufacturers a further incentive to focus on QWERTY, to the exclusion of alternative systems. Positive feedbacks reinforced QWERTY's early lead until it gained virtually the whole market. The superior 'Ideal' keyboard layout, introduced in 1893, appeared too late to disrupt a lock-in to QWERTY.

David (1986) concluded, 'competition in the absence of perfect futures markets drove the industry prematurely into standardization on the wrong system'. Critical to both the emergence and persistence of QWERTY was that the 'larger system of production', comprising typists, employers, manufacturers, and typing instructors, 'was nobody's design'; it was characterized by decentralized decision making. Liebowitz and Margolis (1990) responded, in effect, that 'design' rather than positive feedbacks controlled the process that produced the QWERTY standard. Early typewriter manufacturers, they noted, competed vigorously on features of their machines, and they inferred from this that QWERTY succeeded due to a market test of its relative fitness. Positive feedbacks played no role, they argued, because typewriter suppliers had an opportunity to provide training to offices where they sold their machines. Suppliers could thus internalize, and profit from, the advantages of a superior keyboard.

David (1999, 2001) responded in turn that keyboards were never tested by the market in isolation from numerous other features of machines that varied among manufacturers. Furthermore, Liebowitz and Margolis offered no evidence that typewriter manufacturers found it practical to offer training in touch typing before the 1920s, long after QWERTY had become the established standard. Thus their argument has no empirical basis. Still, David's empirical evidence appears less than conclusive in light of the points raised by his critics.

Liebowitz and Margolis devoted most of their article to matters less relevant to David's argument. They refuted a popular account that QWERTY won 'once and for all' due to the publicity it received when a touch-typing QWERTY typist won a single typing contest in 1888. As they showed, non-QWERTY typists soon won other typing contests, so a single contest could not have been decisive. Their refutation did not, however, address David's argument that the contest in question had publicized the value of touch typing, which was being taught at the time only for QWERTY. Liebowitz and Margolis also refuted the mistaken story that QWERTY had been designed to slow typists down, but, again, this story was never part of the claims made by theorists of path dependence.

Liebowitz and Margolis did convincingly refute one claim about QWERTY that David had tacitly accepted – that the Dvorak Simplified Keyboard, invented in 1932, was so superior to QWERTY that the cost of retraining could be recovered in a period of weeks. As Liebowitz and Margolis showed, this claim was based on dubious experiments, and it does not stand the test of reasoned inference from users' behaviour. But David had mentioned the claim, in a single sentence, only to establish the extent to which the legacy of early events had mattered. His argument is little affected if the relative inefficiency of QWERTY is only on the order of 10%, as estimated by a leading researcher in industrial design and ergonomics (Norman 1990). David's claim about history mattering would, however, be affected if QWERTY's relative inefficiency is next to nothing, as Liebowitz and Margolis suggest.

Videocassette Recorders and Similar Cases

Another influential case study in path dependence was the competition between alternative videocassette recording systems from the mid-1970s to the mid-1980s. The VHS system of JVC (Japan Victor Corporation) became the standard, beating out Sony's Betamax. Arthur (1990) explained this as the result of positive feedbacks in the video rental market, as video stores stocked more film titles for the system that accidentally gained a larger user base, while consumers bought the system for which they could rent more videos. Liebowitz and Margolis (1995) pointed out, however, that Sony had actually been first to market. If positive feedbacks had mattered, they argued, then Sony should have won. They attributed the VHS victory to active product promotion and to the advantage of VHS in offering a longer playing time. In their view, purposeful, forward-looking behaviour had overridden positive feedbacks, ensuring the superior outcome. They offered substantial evidence against Arthur's suggestion that the winning system may have been technically inferior.

The extensively documented account of Cusumano et al. (1992) showed, however, that purposeful behaviour did not trump path dependence. There was indeed a positive-feedback dynamic in the video rental market, but this market emerged late, after VHS had already gained a strong lead. The onset of positive feedbacks turned Betamax's small but stable market share into a fast-declining one, forcing it to exit the market.

More intriguingly, Cusumano, Mylonadis and Rosenbloom attributed the earlier lead of VHS to path dependence in supplier choices. Manufacturers and distributors increasingly supported VHS over Betamax as they saw others doing so, increasing their expectations that VHS, not Betamax, would later become the standard. Ultimately, the authors argued, VHS won as the result of non-systematic differences in the promoters' early strategy choices. First, Sony initially pursued a go-it-alone strategy, while JVC built a coalition of suppliers in order to benefit from positive feedbacks. Second, JVC's partner Matsushita installed a large manufacturing capacity to solidify expectations among other suppliers. Third, Sony opted for a smaller cassette size, while JVC chose a larger cassette with longer playing time. In the event, a longer playing time proved more important to consumers in the early years, when only a VHS tape could record an entire American football game or a long movie. Distributors responded to this temporary advantage by joining the VHS coalition permanently.

This account shows that path dependence is fully compatible with forward-looking behaviour, provided that foresight is imperfect when early choices are made about strategy and product characteristics. Indeed, market participants recognized positive feedbacks, and they sought to influence the early events that would have a disproportionate effect on later outcomes.

Such behaviour is common in advanced-technology industries, and innovators whose forward-looking behaviour takes positive feedbacks into account are more likely to win their markets (Morris and Ferguson 1993; Shapiro and Varian 1998). Indeed, according to many observers, either IBM or Apple Computer rather than Microsoft could have become the dominant firm in microcomputers, controlling the key system standard (Rohlfs 2001; Carlton 1997). However, only Bill Gates of Microsoft had, and acted on, the foresight that control of a standard would matter. He became the world's richest individual as a result.

Such processes are path dependent when outcomes depend, in part, on nonsystematic choices and events. If general foresight is good, however, then systematic considerations may dominate. Market participants may agree on a superior outcome from the start, in the manner of a fulfilled-expectations process (Katz and Shapiro 1985). An example, again, is the CD standard.

What is at stake in a path-dependent process is not necessarily general efficiency or total payoffs. It may, rather, be the distribution of payoffs to different innovators and suppliers. Furthermore, in a path-dependent process, particular individuals can have lasting effects on later outcomes, for better and for worse.

Railway Track Gauge

One individual who made a lasting difference was railway pioneer George Stephenson. Stephenson transferred the gauge of four feet eight and a half inches (1435 mm) from the primitive mining tramways where he gained his early experience to the Liverpool and Manchester Railway. That line became the model of best practice for the earliest railways of Britain, North America, and Continental Europe (Puffert 2000, 2002, 2008). The Stephenson gauge became the standard over wider areas as new railways, interested in compatibility, adopted the gauge of prior neighbouring lines.

Engineers soon came to prefer broader gauges, and they introduced such gauges to new regions. A lack of foresight into the later importance of large-scale network integration led to the emergence of two regional standard gauges in Britain, six in North America, six in Continental Europe, and multiple gauges in Australia, India, and other intercommunicating regions. The cost of coping with or resolving this diversity was the main path-dependent inefficiency in track gauge, outweighing the minor inefficiency of the prevalent Stephenson gauge. Still, diversity was resolved most easily where it proved most costly, and the mechanism for resolving diversity was frequently the sort of coordinating behaviour discussed by Liebowitz and Margolis (1995). Much of Britain's and North America's diversity, for example, was resolved by emerging interregional rail systems that internalized the benefits of standardization.

Even so, these improvements in outcomes were a matter of 'path-constrained amelioration' (David 2001) rather than a complete break from the historical legacy. Britain made the Stephenson gauge its general standard at a time when the consensus of engineers favoured a gauge of five feet to five feet six inches (1524 mm–1676 mm), and North America did so when the consensus favoured five feet. Japan has long regretted its choice of a narrow standard gauge, three feet six inches (1067 mm).

Australia and India have only recently resolved much of their diversity, while the variant gauges of the Iberian peninsula and the former Russian and Soviet empires are becoming more costly as those regions are integrated economically into the core of Europe. However, the cost of this diversity is being reduced by innovative mechanisms that enable trains to change their gauge en route. The potential role of government mandates in improving on path-dependent outcomes was proved in Britain, where the 1846 Gauge Act led to some rationalization of gauges.

Other Railway Standards

Path-dependent diversity in regional standards has also proven costly in such matters as railway electrification systems, clearance dimensions ('loading gauges'), and train control and signalling systems. This diversity has hindered the formation of international high-speed train links in Europe (Puffert 1993).

Several railway standards that were well adapted to early conditions proved poorly adapted to later ones, but they continued in use due to the cost of converting the installed base. Examples reportedly include mechanical couplings and air brakes, as electrical systems would now be safer and less labour-intensive (Hilton 1990, p. 294).

A more famous example is what Veblen (1915) called the 'silly little bobtailed carriages' used in British goods traffic. A long literature has addressed how the historical legacy of interrelated freight handling facilities prevented the modernization of coal cars in particular (Kindleberger 1964, pp. 141–4). Recently, Van Vleck (1997) argued that small coal cars were well adapted to the larger system of distribution, chiefly by reducing the costs of small deliveries. Scott (2001) showed, however, that few coal users benefited from small car-size deliveries. Rather, the cars' small size, widely dispersed ownership (by collieries), antiquated braking and lubrication systems, and generally poor physical condition made them quite inefficient indeed. Replacing these cars and associated infrastructure with modern, larger wagons owned and controlled by the railways would have offered savings in railway operating costs of about 56%, yielding a social rate of return of 24% on the physical costs of conversion. Nevertheless they were not replaced until both the railways and the collieries were nationalized after 1945. Until then, regulations forced the railways to accept colliery cars at set rates or else offer high levels of compensation. Due to technical interrelatedness, the railways could not have saved much in operating costs until virtually all the antiquated cars were replaced, so high transaction costs prevented transition to a more efficient practice.

Further Cases

Cowan (1990) argued that transitory circumstances led to the establishment of the prevalent 'light-water' design for civilian nuclear power reactors. This design, adapted from nuclear submarines, was rushed into use due to the political value of demonstrating peaceful uses for nuclear technology. Thereafter, learning effects arising from engineering experience continued to make the light-water design the rational choice for new reactors. Cowan argued, however, that an equivalent degree of development would likely have made an alternative design superior.

Cowan and Gunby (1996) addressed farmers' choices between systems of chemical pest control and integrated pest management (IPM), which uses predatory insects to devour harmful ones. As the drift of chemical pesticides from neighbouring fields often makes the use of IPM impossible, IPM must be used on the whole set of farms that are in proximity to one another. Where this set is large, the transaction costs of persuading all farmers to forgo chemical methods often prevent adoption. In addition to these localized positive feedbacks, local learning effects also make the choice between systems path dependent. Local lock-in to each practice is sometimes upset by such developments as invasions by new pests and the emergence of resistance to pesticides.

See Also

- ▶ Irreversible Investment
- ▶ Learning-by-Doing
- ▶ Network Goods (Empirical Studies)
- ▶ Network Goods (Theory)
- ▶ Path Dependence
- ▶ Technical Change
- ▶ Veblen, Thorstein Bunde (1857–1929)

Bibliography

Arthur, W.B. 1989. Competing technologies, increasing returns, and lock-in by historical events. *Economic Journal* 99: 116–131.

Arthur, W.B. 1990. Positive feedbacks in the economy. *Scientific American* 262(February): 92–99.

Arthur, W.B. 1994. *Increasing returns and path dependence in the economy.* Ann Arbor: University of Michigan Press.

Carlton, J. 1997. *Apple: The inside story of intrigue, egomania, and business blunders*. New York: Times Business.

Cowan, R. 1990. Nuclear power reactors: A study in technological lock-in. *Journal of Economic History* 50: 541–567.

Cowan, R., and P. Gunby. 1996. Sprayed to death, path dependence, lock-in and pest control strategies. *Economic Journal* 106: 521–542.

Cusumano, M.A., Y. Mylonadis, and R.S. Rosenbloom. 1992. Strategic maneuvering and mass-market dynamics: The triumph of VHS over beta. *Business History Review* 66: 51–94.

David, P.A. 1975. *Technical choice, innovation and economic growth: Essays on American and British experience in the nineteenth century.* Cambridge, UK: Cambridge University Press.

David, P.A. 1985. Clio and the economics of QWERTY. *American Economic Review: Papers and Proceedings* 75: 332–337.

David, P.A. 1986. Understanding the economics of QWERTY: The necessity of history. In *Economic history and the modern economist*, ed. W.N. Parker. Oxford: Oxford University Press.

David, P.A. 1987. Some new standards for the economics of standardization in the information age. In *Economic policy and technological performance*, ed. P. Dasgupta and P. Stoneman. Cambridge, UK: Cambridge University Press.

David, P.A. 1991. The hero and the herd: Reflections on Thomas Edison and the ‘battle of the Systems’. In *Favorites of fortune, technology, growth, and economic development since the industrial revolution*, ed. P. Higonnet, D.S. Landes, and H. Rosovsky. Cambridge, MA: Harvard University Press.

David, P.A. 1997. *Path dependence and the quest for historical economics: One more chorus of the ballad of QWERTY.* Discussion Papers in Economic and Social History No. 20, University of Oxford.

David, P.A. 1999. *At last, a remedy for chronic QWERTY-skepticism!* Working paper, All Souls College, Oxford University.

David, P.A. 2001. Path dependence, its critics and the quest for ‘historical economics’. In *Evolution and path dependence in economic ideas, past and present*, ed. P. Garrouste and S. Ioannides. Cheltenham/Northampton: Edward Elgar.

Frankel, M. 1955. Obsolescence and technological change in a maturing economy. *American Economic Review* 45: 296–319.

Hilton, G.W. 1990. *American narrow-gauge railroads*. Stanford: Stanford University Press.

Katz, M.L., and C. Shapiro. 1985. Network externalities, competition, and compatibility. *American Economic Review* 75: 424–440.

Kindleberger, C.P. 1964. *Economic growth in France and Britain, 1851–1950*. Cambridge, MA: Harvard University Press.

Liebowitz, S.J., and S.E. Margolis. 1990. The fable of the keys. *Journal of Law and Economics* 33: 1–25.

Liebowitz, S.J., and S.E. Margolis. 1995. Path dependence, lock-in, and history. *Journal of Law, Economics, and Organization* 11: 204–226.

Morris, C.R., and C.H. Ferguson. 1993. How architecture wins technology wars. *Harvard Business Review* 71-(March–April): 86–96.

Nelson, R.R., and S.G. Winter. 1977. In search of a useful theory of innovation. *Research Policy* 6: 36–76.

Norman, D.A. 1990. *The design of everyday things*. New York: Doubleday.

Puffert, D.J. 1993. Technical diversity and the integration of the European highspeed train network. In *High-speed trains, fast tracks to the future*, ed. J. Whitelegg, S. Hultén, and T. Flink. Hawes: Leading Edge.

Puffert, D.J. 2000. The standardization of track gauge on North American railways, 1830–1890. *Journal of Economic History* 60: 933–960.

Puffert, D.J. 2002. Path dependence in spatial networks: The standardization of railway track gauge. *Explorations in Economic History* 39: 282–314.

Puffert, D.J. 2004. Path dependence, network form, and technological change. In *History matters: Essays on economic growth, technology, and demographic change*, ed. T. Guinnane, W. Sundstrom, and W. Whatley. Stanford: Stanford University Press.

Puffert, D.J. 2008. *Tracks across continents, paths through history: The economic dynamics of standardization in railway gauge*. Chicago: University of Chicago Press.

Rohlfs, J.H. 2001. *Bandwagon effects in high-technology industries*. Cambridge, MA/London: MIT Press.

Scott, P. 2001. Path dependence and Britain’s ‘coal wagon problem’. *Explorations in Economic History* 38: 366–385.

Shapiro, C., and H.R. Varian. 1998. *Information rules*. Cambridge, MA: Harvard Business School Press.

Van Vleck, V.N.L. 1997. Delivering coal by road and rail in Britain: The efficiency of the ‘silly little bobtailed’ coal wagons. *Journal of Economic History* 57: 139–160.

Veblen, T. 1915. *Imperial Germany and the industrial revolution*. London: Macmillan.

Patinkin, Don (1922–1955)

Nissan Liviatan

Abstract

Don Patinkin’s main contributions are in monetary theory, including the topics of involuntary unemployment and the interpretation of the writings of J.M. Keynes. He criticized the classical and neoclassical monetary model for its ‘invalid dichotomy’ between the real and the monetary sectors. His main underlying concern was whether capitalism possessed an automatic mechanism for attaining full employment. He claimed that the real interest rate might be insufficiently flexible and the real

balance effect insufficiently powerful to allow any rapid convergence to equilibrium, rendering it politically unrealistic to rely on automatic forces to establish full-employment equilibrium in reasonable time.

Keywords
Capitalism; Chicago School; Clower constraint; Cowles Commission; Friedman, F.; Involuntary unemployment; IS–LM model; Keynes, J.M.; Kuznets, S.; Monetary economics, history of; Neoclassical monetary theory; Patinkin, D.; Quantity theory of money; Real vs monetary sector dichotomy; Real-balance effect; Samuelson, P.; Stability analysis; Walras's Law

JEL Classifications
B31

Don Patinkin is regarded as the 'father of the economics profession' in Israel. Upon his arrival in Israel with his wife Dvora in 1949, he joined the Hebrew University and raised a generation of students trained in modern economics (known as the 'Patinkin boys'), who were to form the backbone of the economics departments in the various universities, the staff of Treasury and the Bank of Israel, the commercial banks and the other institutions that had a demand for economists. In spite of his young age, he was an economist with outstanding academic achievements, which marked him as a rising star in the economics profession. His choice to live in Israel was a source of much pride to the new state. He passed away in 1995, but the impact of his teaching and of his personality will linger on for many years.

The Chicago Days

Don Patinkin was born in 1922 in Chicago to an Orthodox Jewish family that lived in a predominantly Jewish neighbourhood. His early education was a combination of secular and rabbinical studies. In 1943 he enrolled in the economics department of the University of Chicago, obtaining his Ph.D. in 1947. This period left a deep impression on Patinkin and had a great impact on his future work. He was influenced not only by the Chicago Tradition of free-market liberalism, but also by the personalities of his prominent teachers Frank H. Knight, Jacob Viner, Henry C. Simons, Oscar Lange and Lloyd W. Mints. (See his account of his teachers in Patinkin, 1981b.)

In addition, he was awarded a fellowship with the Cowles Commission, then situated in Chicago, which hosted a remarkable number of prominent economists. Patinkin looked back to his Chicago days with pleasure and nostalgia, and considered himself lucky to have benefited from the contact with such 'giants'. His Ph.D. dissertation, 'On the consistency of economic models: a theory of involuntary unemployment', under the supervision of Jacob Marshak (the chairman of the Cowles Commission), is on a topic to which he returned many times over the years without being able to find a satisfactory solution (nor has any other economist).

The Years at the Hebrew University

In 1949 he accepted the proposal of the Hebrew University of Jerusalem to serve as a senior lecturer in the economics department. On his very first day Patinkin plunged into this task, directing the transition of the department from the Continental descriptive and institutional framework to the Anglo-Saxon tradition of analytical economics (Barkai, 1993). Professor Alfred Bonne, who chaired the traditional economics department, supported this move. In the first years he taught practically all the courses in microeconomics and macroeconomics, at all levels, and performed this task outstandingly.

It is remarkable that these years of great pressure were also the most fruitful of his career: he completed his monumental book *Money, Interest and Prices* (MIP, 1956) and wrote a number of influential papers in leading journals, usually rebutting criticisms on various topics related to the book (such as the invalid dichotomy, discussed later). In evaluating the numerous reviews of the book, Stanley Fischer (1993) states that practically all the reviewers recognized that they were dealing with a major work.

To build the foundations of the economics profession in Israel, Patinkin sent a group of graduates, whom he considered candidates for an academic career, for Ph.D. studies in the top universities in Britain and the United States. The building of foundations included also the construction of a statistical base of the Israeli economy. To perform this task he was appointed director of the Falk Institute of Economic Research Israel in 1956, where he continued the work of Daniel Creamer and Harold Lubell, the previous directors. It seems that Simon Kuznets, who was involved in formulating the programme of the Falk Institute, had a profound influence on Patinkin's interest in empirical research. In addition to directing numerous research projects at the institute, Patinkin himself wrote on the early years of the Israeli economy (*The Israeli Economy in the First Decade*, 1959a), where the elimination of the monetary overhang associated with repressed inflation fitted well with his model of the real balance effect.

Although he believed in, and represented, the Chicago pro-market creed, he never pushed this approach forcefully. The very first lesson in his celebrated course 'Introductory Economics', modelled on the famous textbook of Samuelson with application to Israel, was about the allocation of scarce resources among competing uses, which could in principle be performed by the market or by a central planning committee.

Patinkin completed his term of office as chairman of the department of economics in 1960, moving on to serve as the Dean of Social Sciences, and from there in 1980 to serve as Rector and finally as President of the Hebrew University. In all these years he maintained his touch with monetary economics, especially from the doctrinal aspect.

Patinkin participated actively in the debates concerning Israel's economic policy problems. In particular, he was critical of the way monetary policy was run. He served on a number of policy committees (for a thorough discussion of this aspect of Patinkin's activity, see Barkai, 1993) and contributed to the daily press of the early 1970s when the inflationary process began. In later years he preferred that the economists that he had raised should handle these matters.

On the occasion of Patinkin's retirement his colleagues organized a conference in his honour. The scientific works of the participants, who included many of the economists that he regarded highly, were published in *Monetary Theory and Thought* (Barkai et al. 1993), which covered topics related to Patinkin's work.

Patinkin's Contribution to Monetary Economics

Patinkin contributed to three main areas in economic theory: his criticism of neoclassical monetary theory, his treatment of involuntary unemployment and his work on the history of economic thought, in particular the writings of Keynes. Patinkin introduced some order into the vague (some may prefer the term 'chaotic') state of the monetary model that existed in his time. MIP stands out as a bridge between pre-Keynesian economics, Keynesian economics and the modern economic literature. Its economic rigour, building the macroeconomic model on micro foundations, was unprecedented in the literature on monetary economics.

Patinkin was very critical of the monetary model formulated by the classical and neoclassical theorists. In particular, he claimed that their theory was 'guilty' of the 'invalid dichotomy' between the determination of relative prices and the absolute price level. More specifically, the dichotomy relates to the separation between the real sector, where relative prices are determined, and the monetary sector, where the absolute price level is determined by some version of the quantity theory of money (the Cambridge equation). He claimed that this dichotomization is invalid because, by Walras's Law, the excess demand for money is just the sum of excess supplies in all other markets and hence must share the same parameters, in particular the money supply. The fact that in the neoclassical formulation the money supply appears only in the money market is self-contradictory. (In the second edition of MIP, 1965, Appendix to ch. 8, Patinkin pointed out that, when the real balance effect is confined to the bond market, it is possible to express the excess demand functions for commodities in terms of relative prices and the interest rate without referring explicitly to the real balance effect.).

To prove that the neoclassical monetary economists adhered in fact to the invalid dichotomy, Patinkin created a ‘database’ of the relevant writings of these economists (summarized in the first and second editions of MIP), and scrutinized carefully the suspect sentences to show that they were unclear and even reckless. There is no doubt that Patinkin was a master of the literature on monetary theory, and he used it effectively to support his arguments.

Since many people wrote without a formal analytical apparatus in those days, they often said contradictary things concerning the dichotomy, and Patinkin identified and stressed the inconsistencies. The mathematically inclined economists used the formulation of excess demand functions in terms of all the n prices ($p_1,\ldots,p_n$), which can be multiplied by a Lagrange multiplier λ, which could be any positive number. However, in order to reflect the fundamental property of zero degree homogeneity of real excess demand functions with respect to money prices and the nominal money supply, λ has to be set equal to 1/M, where M is the nominal money supply; then it would represent the real balance effect. But Patinkin insisted (and documented) that as a rule they thought of λ as $1/p_n$, that is, they thought of excess demand for commodities as dependent only on relative prices, without taking account of the real balance effect.

The preoccupation with the question of ‘what people really thought’ left a gray area of possible interpretations, which depended on subjective evaluations. Paul Samuelson (1968), who thought that in principle Patinkin’s criticism was well taken, nevertheless believed that Patinkin’s reading of the earlier theorists was not sympathetic. However, the examples of the articles of Hickman (1950) and Archibald and Lipsey (1958), who tried to defend the invalid dichotomy, made it clear that Patinkin’s tough criticism was justified from the point of view of improving professional rigour in economic science (see also Fischer, 1993).

Patinkin’s critical evaluation reflects the stringent criteria he applied to the work of his predecessors. He required of monetarist theorists who put money in the utility function to state explicitly the rationale for holding money; he insisted on an explicit reference to the real-balance effect, and he required an understanding of the difference between the individual and market experiments. In addition, he insisted on the incorporation of stability analysis of the money market in the same way as his predecessors analysed the stability of markets for ordinary commodities. He considered the fulfillment of all these criteria necessary for a full integration of money and value theory.

The ‘victims’ of this harsh criticism included such famous names as Walras, Fisher, Pigou and Cassel, in whose writings the presence of the invalid dichotomy was ‘highly probable’, as well as others who were more explicit about it, such as Lange, Modigliani, and Hickman (Patinkin, 1965, p. 175, n. 33).

Patinkin enjoyed the role of critical interpreter of texts, which he attributed to his training at the Yeshiva College in Chicago (1994). This perhaps explains his infatuation with Keynes’s writings in later years, and his preoccupation with the writings in the Chicago tradition. The former involved mainly the evolution of Keynes’s thoughts on effective demand and involuntary unemployment, and the latter focused on the interpretation of the quantity theory of money and the economic philosophy of his famous teachers at the University of Chicago.

The non-technical writings of Keynes (1936) were a fertile ground for interpretations and formulations of formal models attributed to his ideas, and it provided Patinkin with ample room for clarification of Keynes’s arguments. For example, he presented a diagrammatic exposition of the Keynesian theory in Patinkin (1982), especially Figures 5 and 6, clarifying the concepts of effective demand and aggregate supply in the Keynesian model. (In Figure 6 it is shown that effective demand is determined at the intersection of aggregate demand and supply – in terms of wage units – as functions of employment. In this diagram the real wage is endogenous to the level of employment on the assumption that firms are on the demand curve for labour. Thus the real wage is indirectly determined by aggregate demand. In this sense it is not a fixed-price model.) In his analysis of Milton Friedman's statement of the quantity theory of money (Patinkin, 1981) he contrasts it critically with what Patinkin considered the true Chicago tradition.

Involuntary Unemployment

While the task of putting the house of neoclassical monetary theory in order involved an in-depth analysis of the early literature, his other major preoccupation was in an area which required his own creativity – involuntary unemployment. This problem, which reflected the realities of the Great Depression of the 1930s, occupied Patinkin's academic interests from his Ph.D. dissertation and throughout his later work. Yet the problem of why the workers could not avoid unemployment by real wage cuts remains basically unresolved to this very day.

Patinkin first approached this problem in his famous early article in the *American Economic Review* (1948), where he claimed that the real interest rate and the real-balance effect might not be sufficiently flexible to allow an equilibrium solution, and even if they did it may take a long time (due to bankruptcies and pessimistic expectations). This may render it politically unrealistic to rely on automatic forces to establish full-employment equilibrium. Patinkin therefore considered unemployment essentially in the context of economic dynamics.

In Chapter 13 of MIP, Patinkin took an additional step in dealing with this issue, arguing that if firms cannot sell their optimal (competitive) output they will not employ their optimal labour input. This gave rise to a new area of research in macroeconomics, namely, disequilibrium models. Barro and Grossman (1971) combined this analysis with the Clower constraint, which postulates (as explained by Barro and Grossman) that, if workers cannot supply their optimal labour services they will not purchase their optimal (competitive) quantity of goods. Barro and Grossman go on to show how equilibrium can be established in the fixed-price model of this type. Over the years, the criticism of these models increased because they required arbitrary rationing rules (Drazen, 1980), and because they were too complicated technically. The disappearance of widespread involuntary unemployment in the post-Second World War era probably had something to do with the growing unpopularity of these models.

It is noteworthy that Patinkin refrained (in the second edition of MIP) from seeking a solution to the problem of involuntary unemployment in the domain of imperfect competition, in spite of Arrow's (1959) remark that in disequilibrium situations the competitive model is problematic. It seems that this is an indication of Patinkin's conservative approach to economic analysis.

Although most of MIP is devoted to the working of Patinkin's model in full employment, the more interesting implications of monetary policy were in connection with unemployment. The latter case gave rise to the fundamental question of whether the capitalist system possesses an *automatic* mechanism for attaining full employment, which is the basic problem that underlies much of Patinkin's work. Perhaps this explains why he was willing to take the risk of dealing with disequilibrium models, although he realized their limitations (1965, ch. 13, n. 9).

Some of the issues which were presented in MIP gave rise to criticism by prominent economists. But in all these confrontations Patinkin had the upper hand. One can cite as an example Hicks's (1957) criticism of Patinkin's interpretation of Keynesian unemployment theory; Patinkin's reply (1959b) in terms of the Hicksian IS–LM model suggested that Hicks did not fully understand Pigou's (1943) mechanism of the real-balance effect.

Patinkin's early work dealt solely with the static economy, while the profession was concerned in the 1960s with models of economic growth, including monetary growth. This led Patinkin to write a paper, with David Levhari (1968), on monetary growth in the fashion of Tobin's original contribution to these models.

Patinkin's own view of his early work and his critical reflections about the recent developments in economics are interesting. We have a glimpse of these in the introduction to his final, abridged edition of MIP in 1989, 23 years after the publication of the first edition. In this introduction he welcomes the progress that has been made in disequilibrium theory, although he realizes its limitations, since it contradicts some of the tenets of rational expectations. He also welcomes the renewed theoretical work by the neo-Keynesian economists on the rational basis of price and wage rigidities, and discusses the effect of the new developments related to rational expectations. His discussion is certainly very scholarly but short of the original insights that characterized his earlier writings. It seems that rational expectations represent a whole new philosophy that was absent in the writing of the 1950s and 1960s, which one might call the age of innocence.

See Also

- ► Keynes, John Maynard (1883–1946)
- ► Monetary economics, history of.

Selected Works

1948. Price flexibility and full employment. *American Economic Review* 37: 543–564.

1956. *Money, interest and prices*. Evanston, IL: Row Peterson.

1959a. *The Israeli economy – the first decade*. Jerusalem, The Maurice Falk Institute For Economic Research in Israel.

1959b. Keynesian economics rehabilitated: A rejoinder to Professor Hicks. *Economic Journal* 69: 582–587.

1965. *Money, interest and prices*, 2nd edn. New York: Harper and Row.

1968. (With D. Levhari.)The role of money in a simple growth model. *American Economic Review* 58: 713–753.

1981a. The Chicago tradition, the quantity theory, and Friedman. In *Essays in and on the Chicago tradition*. Durham, NC: Duke University Press.

1981b. Introduction: Reminiscences of Chicago 1941–47. In *Essays In and On the Chicago Tradition*. Durham, NC. Duke University Press.

1982. A critique of Keynes' theory of effective demand. In *Anticipations of the general theory, and other essays on keynes*. Chicago: University of Chicago Press.

1989. *Money, interest and prices*, 2nd edn. Abridged, with a new introduction. Cambridge, MA: MIT Press.

1994. From Chicago to Jerusalem. *Economic Quarterly*, Hebrew, New Series 1994(2): 165–190.

I am grateful to Akiva Offenbacher and Stanley Fischer for their comments. I also benefited greatly from the comments and suggestions of my veteran colleagues of the economics department of the Hebrew University: Haim Barkai, Nachum Gross, Michael Michaely, Ephraim Kleiman and Gur Ofer

Bibliography

Archibald, G., and R. Lipsey. 1958. Monetary and value theory: A critique of Lange and Patinkin. *Review of Economic Studies* 26: 1–22.

Arrow, K. 1959. Towards a theory of price adjustment. In *The allocation of economic resources*, ed. M. Abramovitz et al. Stanford, CA: Stanford University Press.

Barkai, H. 1993. Don Patinkin's contribution to economics in Israel. In *Monetary theory and thought*, ed. H. Barkai, S. Fischer, and N. Liviatan. London: Macmillan.

Barro, R., and H. Grossman. 1971. A general disequilibrium model of income and employment. *American Economic Review* 61: 82–93.

Drazen, A. 1980. Recent developments in macroeconomic disequilibrium theory. *Econometrica* 48: 283–306.

Fischer, S. 1993. Money, interest and prices. In *Monetary theory and thought*, ed. H. Barkai, S. Fischer, and N. Liviatan. London: Macmillan.

Hickman, W. 1950. The Determinacy of absolute prices in classical economic theory. *Econometrica* 18: 9–20.

Hicks, J. 1957. A rehabilitation of 'classical economics'? *Economic Journal* 47: 278–289.

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

Pigou, A. 1943. The classical stationary state. *Economic Journal* 53: 343–351.

Samuelson, P. 1968. What classical and neoclassical monetary theory really was. *Canadian Journal of Economics* 1: 1–15.

Patten, Simon Nelson (1852–1922)

A. W. Coats

Keywords

American Economic Association; Patten, S. N.; Protection; James, E. J.

JEL Classifications

B31

One of the most original and idiosyncratic American economists of his generation, Patten was born at Sandwich, Illinois on 1 May 1852 and studied at Jennings Seminary, Aurora, Illinois. There he met Joseph French Johnson, later a colleague at the University of Pennsylvania, whom he followed to Halle in 1876 after spending only 18 months as a freshman at Northwestern University. At Halle Patten obtained the Ph.D. degree remarkably quickly, in 1878, and he encountered two major personal influences, his teacher Johannes Conrad and a fellow American student, Edward Janes James, who was eventually instrumental in securing Patten's appointment at the University of Pennsylvania in 1888, where he remained throughout his academic career. In the intervening period, however, like Thorstein

Veblen, Patten had been unable to get a university post despite the publication of his highly original *Premises of Political Economy* (1885), and was obliged to work on a farm and teach in various public schools, partly because of his poor eyesight.

Once at Philadelphia, Patten proved to be a profoundly stimulating pedagogue and author of a series of unusual, even eccentric books that challenged, provoked and sometimes baffled his professional peers. In harmony with the Wharton School tradition, he was an ardent protectionist, believing that trade barriers would stave off the dangers envisaged by Ricardo and Malthus. Adopting an optimistic, teleological view of the prospects for American abundance, provided that crop variations could be developed to counteract soil exhaustion, Patten insisted that economic laws were not natural, but social. His conception of economics was broad, as in the German tradition, yet his own work was abstract and deductive rather than heavily empirical or statistical. Together with James, he tried in 1884 to form a Society for the Study of National Economy, modelled on Conrad's suggestions, but when this failed to gain sufficient support they joined Ely and others in launching the American Economic Association, of which Patten was elected president in (1908–9). Patten's concepts of the laws of pleasure and pain, his theory of consumption, and his idea of the social surplus were intriguing but puzzlingly novel and unsystematic, yet his awareness of the costs of growth and his concern for the environment anticipated late 20th-century anxieties.

Selected Works

1885. *The premises of political economy: Being a re-examination of certain fundamental principles of economic science*. Philadelphia: J.B. Lippincott.

1889. *The consumption of wealth*. Philadelphia: T. & J.W. Johnson. 2nd ed. Philadelphia: published for the University; Boston: Ginn & Co., 1901.

1890a. *The economic basis of protection*. Philadelphia: J.B. Lippincott.

1890b. *The educational value of political economy.* Baltimore: American Economic Association.

1896. *The theory of social forces*. Philadelphia: American Academy of Political and Social Science.

1899. *The development of English thought: A study in the economic interpretation of history.* New York/London: Macmillan.

1902. *The theory of prosperity.* New York/London: Macmillan.

1924. *Essays in economic theory*, ed. R.G. Tugwell. New York: A.A. Knopf.

Bibliography

Fox, D.M. 1967. *The discovery of abundance: Simon N. Patten and the transformation of social theory*. Ithaca: Cornell University Press.

Payment Systems

William Roberds

Abstract

Payment systems are arrangements that allow for the discharging of debts by the transfer of specialized claims. This article illustrates how payment systems can facilitate exchange in economic environments where enforcement of obligations is limited, and collateral is scarce.

Keywords

Banking industry; Banknotes; Bills of exchange; Central banking; Fiat money; Inside money; Intermediate goods; Net settlement; Network effects; Outside money; Payment systems; Settlement; Wicksell triangle

JEL Classifications

E42; E44

A *payment* occurs when one party, the *payer*, transfers an asset to another party, the *payee*, for the purpose of discharging a debt incurred by the payer. Or, a payment may consist of the payer's instruction to a third party to make such a transfer, as is the case with a cheque payment. While in principle a payment may be made with any asset, in practice virtually all modern payments involve transfers of debt claims on either central banks (including 'outside money' in the form of both currency and deposits) or private banks ('inside money', today almost always in the form of deposits). Available evidence suggests that most payments are still made in cash, but these transactions tend to be for relatively small amounts. By value, the wide majority of payments involve transfer of bank deposits by various means.

A payment may or may not constitute *settlement*, a legal discharge of a debt. In most countries, for example, a payment by means of a transfer of claims on a central bank unconditionally settles a debt, whereas other types of payment settle a debt only after certain conditions have been fulfilled (for example, after a cheque has been honoured by the bank on which it is drawn).

A payment *system* is a collection of technologies, laws, and contracts that allow payments to occur and determine when a payment effects a settlement. Payment systems include currency, cheques, credit and debit cards, electronic funds transfers, and so on. Developed economies depend critically on the near-flawless operation of such systems. By offering debtors low-cost and trustworthy means of settling their debts, payment systems provide an important stimulus to the use of credit, and to economic activity more generally.

Some simple statistics illustrate these assertions: in the year 2003, 81 billion payments of $824 trillion were recorded in the United States, not counting payments made in currency (Committee on Payment and Settlement Systems 2005). Another way of framing these numbers is to note that they imply, on average, $75 in non-cash payments for each dollar of final output produced in the United States in 2003. During the same year each US resident made 278 non-cash payments on average. All developed economies display similar levels of payments activity.

Theory of Payments

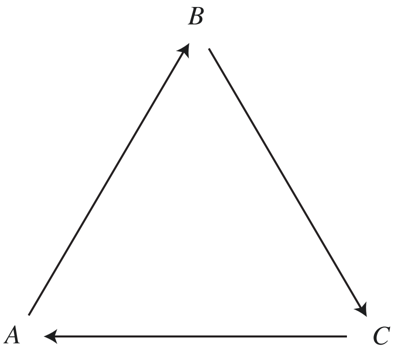
Despite their ubiquity and their obviously central role in modern economies, payments have only recently begun to make their way into mainstream economic theory. Payment systems do not exist in Arrow–Debreu economies, where transfers may always be made in kind, and promises to transfer are enforced by a social planner. In these economies there is no need for specialized assets to allow for payments, technologies for transferring these assets, or rules concerning when such transfers settle a debt.

Even if the planner's ability to enforce promises is limited, payments may still be inessential. Agents will have incentives to honour their obligations so long as they have access to sufficient amounts of collateral that can be attached by creditors after a default. Payment systems become relevant when enforcement is limited and collateral is scarce. In such environments, payment systems serve as devices that allow for enforcement of debts while making efficient use of available collateral.

One commonly available type of collateral is, of course, outside fiat money, but a discussion of the comparative payment roles of inside and outside money is beyond the scope of this essay. Two influential papers in this area have been Freeman (1996, see especially its discussion in Green 1999) and Cavalcanti and Wallace (1999). For the reminder of this article I will concentrate on payments in private debt.

An Illustration

To demonstrate the function of payment systems, I consider some models of payment based on the celebrated 'Wicksell triangle' depicted in Fig. 1. Each of the three agents is endowed with a unit of a generic numeraire good. Agent A has the possibility of converting this good into a 'customized' good that is (highly) desired by agent B, who can convert his numeraire into a good desired by agent C, who can produce a good that is desired by A.



Payment Systems, Fig. 1 Wicksell triangle

Barring difficulties in enforcement, efficiency would require each agent to produce the appropriate customized good and deliver it to the next agent. I call this allocation the *full-enforcement efficient* allocation.

The general goal of payment systems is to deliver an allocation that approximates this allocation, to the extent this is feasible under limited enforcement. I now consider to what extent various payment systems are able to do this. In each of these environments, any enforcement actions will occur through a fourth agent known as the *centre* or 'central counterparty', who has a restricted ability to punish agents who default on their obligations. Punishments may include limited fines, attachment of collateral, and public announcements of a default.

Payment Model 1: 'Netting'

Kahn et al. (2003) analyse the following version of the Wicksell-triangle environment. A, B, and C each consists of a buyer-seller pair who live at a separate 'location', meaning that trade occurs as bilateral encounters between buyer and seller. Agents are not particularly inclined to keep their promises, but may post some numeraire as collateral before trading begins. There is a single period during which sellers can visit buyers and transfers of customized goods can occur, and a subsequent period during which numeraire may be transferred. Prices of goods are given in numeraire and are determined through bilateral negotiations.

In this environment, it is easy to show that the amount of collateral required for trade can be minimized by the use of a payment system based on *net settlement.* However, as net settlement typically requires the diversion of resources in order to acquire and post costly collateral, its use will entail a welfare loss, relative to the full-enforcement efficient allocation.

Under net settlement, after trades have occurred, the central counterparty sums for each agent the amount the agent owes *to* the seller he bought from, minus the amount owed *from* the agent from the buyer sold to. If this sum is positive, the agent transfers numeraire to the central counterparty, and if the amount is negative, he receives numeraire from the central counterparty.

Payment in this environment simply consists of an agent's declaration of his intent to settle, and this occurs simultaneous with trade. *Settlement* is the two-stage process of (*a*) replacing gross obligations with net obligations and (*b*) discharging net obligations through transfer of numeraire.

A characteristic feature of net settlement is 'set-off,' under which a debt owed *by* party *X* is enforced by cancelling it ('setting it off') against its debt owed *to* party *X*. In this fashion, agent *X*'s creditor may exercise a de facto prior claim against *X*, even when other means of exercising priority are costly (such as posting additional collateral). Payment systems incorporating net settlement allow set-off to occur in a regular and predictable fashion.

Netting of obligations is an ancient method of payment, dating at least to the 13th century fairs of Champagne (Kohn 2001). It continues to be used extensively for settling high-value, recurring obligations such as those that arise between commercial banks (for example, the CHIPS system which operates in the United States). But there are certain limitations that prevent its more widespread use. The first is that there may be an inadequate legal basis for netting (Bliss 2003). Second, netting works well only if all parties involved are of roughly equal creditworthiness (Kahn and Roberds 2003). Finally, netting may require too much coordination in the sense that all parties must agree in advance to participate in the netting arrangement. These limitations have given rise to other forms of payment systems which, in effect, allow netting to occur in a more decentralized fashion.

Payment Model 2: 'Banknote'

Kiyotaki and Moore (2000) discuss a slightly different model from model 1 above. Suppose that preferences and endowments are the same as above, but that bilateral encounters between agents are separated in time: agent *C* first has an opportunity to buy his desired good from agent *B*, who then has an opportunity to buy from *A*, who can then buy from agent *C*. Agent *C* is known to be creditworthy but *A* and *B* are not.

In this model, the full-enforcement efficient allocation can be implemented if *C*'s debt can 'circulate'. More specifically, *B* receives debt from *C* in return for a customized good. Agent *B* then trades *C*'s debt to *A*, in return for *A*'s customized good. *A* then presents *C*'s debt to *C* for redemption. Finally, agent *C* completes the cycle of trade by transferring a customized good to *A*. *Payment* in this environment corresponds to either the issue (by *C)* or transfer (by *B*) of *C'*s debt. If *C* is sufficiently creditworthy, *B*'s transfer of *C*'s debt will also constitute a *settlement.* Otherwise settlement may not occur until *C* redeems his debt.

Under this arrangement it is not necessary for all parties to be creditworthy for trade to occur. Agent *C* may enjoy some natural advantage in this regard. This advantage could take the form of ownership of attachable assets or, in a dynamic setting, it could be that people have better information on the actions of *C* than on the actions of other agents (Cavalcanti and Wallace 1999). In this arrangement, *C*'s debt becomes a form of specialized asset for use in payment, a 'banknote'.

This is the basic model for many transactions using not only privately issued banknotes (which are rarely observed nowadays) but also other means of transferring debt claims. A retail store may not be willing to accept a customer's IOU in exchange for merchandise but is perfectly willing to accept a debt (that is, deposit) claim on a bank, transferred by means of a credit or debit card.

This form of payment also has a long history. One of the most famous early examples is from 15th-century Genoa. There, payments were commonly made using claims on an institution responsible for managing the debt of the state (the *Casa di San Giorgio*; see Kohn 1999). Under this arrangement agent *C* became, in effect, an agent of the state, whose creditworthiness derived from the taxation powers delegated to it. People owing taxes could use claims on the *Casa di San Giorgio* to discharge their own tax obligations, which generated a demand for these claims as payment instruments.

Note that model 2, like model 1, involves a form of netting. When a consumer purchases merchandise with, say, a debit card, the consumer is in effect netting out the debt he owes *to* the merchant against debt (deposits) owed him *by* his bank. In contrast to model 1, however, there need be no prior agreement between merchant and consumer, given sufficient trust in the banking system.

Payment Model 3: 'Bank Loan'

Model 2 illustrates how payment systems allow netting to occur in a decentralized fashion. This model is inadequate for some situations, however, because it does not explain the simultaneous existence of both liquid and illiquid debt. In particular, this model is inappropriate for production economies where a producer may require prompt delivery of an intermediate good now in order to produce a final good that can be sold only later. In such situations, working capital is typically provided by the issue of debt.

To remedy this shortcoming, some studies have attempted to modify model 2 in order to incorporate both transferable ('liquid') and non-transferable ('illiquid') debt. Kiyotaki and Moore (2000) consider a model which maps into the following variation. Suppose that the timing of the first two transactions in the Wicksell triangle is reversed, so that that agent *B* first has an opportunity to buy from *A*, then *C* from *B*, and finally *A* from *C*. This timing is natural if *B* uses *A*'s good as an intermediate good.

As in model 2, agent *C* is trustworthy but agents *A* and *B* may not be. In addition, Agent *C* enjoys a special privilege as a creditor, that is, an enhanced ability to enforce debts, and serves as 'banker' to agent *B*.

In this modified example, it is possible to show that the full-enforcement efficient allocation can sometimes be implemented through use of a combination of transferable and non-transferable debt. Specifically, suppose that *B* has an opportunity to meet with *C* before production of specialized goods can occur, and before trading begins. Agent *B* issues debt to *C*, and *C* in turn issues debt to *B*. When *B* then encounters *A*, he pays for *A*'s specialized good by transferring *C*'s debt to *A*. Agent *B* then has the opportunity to discharge his debt to agent *C* by transferring his specialized good to *C*. Finally, agent *A* presents *C* with his debt, and receives *C*'s specialized good. *Payment* and *settlement* are defined as in model 2.

In short, in this model agent *C* is engaged in 'liquidity transformation', which consists of holding *B*'s debt, which would be unenforceable by *A*, while issuing to *B* his own enforceable and therefore transferable debt. In practice, this liquidity transformation is usually provided by banks. This function of banks was already well established by the 14th century (Kohn 2001).

Payment Model 4: 'Bill of Exchange'

Model 3 allows for the coexistence of liquid and illiquid debt, but may not be appropriate for all circumstances. In some environments, there may be no agents with special enforcement abilities, such as agent *C* above. This is particularly true for economies with less developed legal and financial systems. Yet through the process of payment it may still be possible to economize on resources devoted to enforcement, by allowing for the discharge of one debt by the transfer of another.

Kahn and Roberds (2001) consider the following variation on model 2. The order of meetings is *A* with B, *B* with C, and *C* with *A*. The customized good produced by agent *C* is now valued by both *A* and *B*.

The full-enforcement efficient allocation can then be supported as follows. Suppose that agent *B* issues debt to *A* in the first transaction, and that agent *C* issues debt to *B* in the second transaction, which is subsequently passed to *A*. In the final transaction, *A* presents *C*'s debt to *C*, and *C* redeems his debt by providing the appropriate good to *A*. *Payment* in this environment again corresponds to the passing of *C*'s debt by *B* to *A*, and *settlement* occurs either simultaneously with payment, or when *C* redeems his debt.

The intuition behind the efficiency of this arrangement is as follows. Suppose that, instead of making use of transferable debt, trade is organized as a 'credit chain' (Kiyotaki and Moore 1997), in which *B* issues debt to A, C issues debt to B, and B promises to discharge his debt with A once he has collected from C. If enforcement is less than perfect and *B* also values C's customized good, then *B* may collect C's debt then 'take the money and run', that is, abscond with C's good. But if *A* requires an 'early' payment from *B* in the form of a transfer of *C*'s debt, *B*'s default can be averted, provided that *A* can respond to a failure to pay at this stage by preventing *B* from collecting with *C*.

As in the models above, enforcement of *B*'s obligation to *A* occurs through a form of netting. By requiring early payment from *B* in the form of *C*'s transferable debt, *A* is in effect forcing *B* to cancel one debt with another. The key distinction between model 4 and earlier models is that this cancellation is no longer instantaneous. In other words, even potentially bad credits such as *B* are allowed to issue debts as long as they agree to punctually pay them off using the debt of another, possibly stronger credit.

The work of economic historians (see Ashtor 1972) suggests that model 4 is also an ancient one. Its use in the West (in the form of bills of exchange and similar instruments) dates from the late 12th century, and likely arose from even earlier Middle Eastern precedents. Even in today's advanced economies, this model persists in the form of trade credit that is granted with the understanding it will be repaid in another form of debt, nowadays typically bank funds.

Payments and Networks

Payment systems based on the models discussed above have been in use for some time. Successful application of these models, however, requires some information which may not always be present in practice. At a minimum, participants in these arrangements must be able to distinguish the identity of their counterparties, and have some notion of their counterparties' ability to honour their debts. Historically, these requirements have often worked to limit the use of many forms of non-cash payments to established businesses, wealthy individuals, or parties already well known to each other.

These constraints have become less onerous with improvements in information technology. In particular, the years since 1960 have seen rapid development of electronic payment systems based on the use of cards (Evans and Schmalensee 1999). A noteworthy distinction between electronic systems and their paper-based counterparts is that the new systems require the use of specialized communications networks.

As is the case with other industries, the presence of 'network effects' in payment systems leads to complications (see Weinberg 1997). Baxter (1983) was the first to point out the essentially 'two-sided' nature of the service provided by these networks: that is, that efficiency in these networks may depend critically on the allocation of their costs between buyers and sellers. This insight has been subsequently expanded on by many authors (an authoritative survey is given in Rochet and Tirole 2004). Nonetheless, as of this writing, no consensus has emerged concerning efficient allocation of services provided by these systems (Evans and Schmalensee 2005).

Conclusion

Payment systems are an important component of decentralized exchange. This article has illustrated how the fundamental role of these systems is the reduction of chains of obligations to a smaller and more readily enforceable set of obligations. Ongoing improvements in information technology have the potential to increase the scope and efficiency of payment systems, and this will require economists to provide more precise models of their function and essential nature.

See Also

- ▶ Banking Industry
- ▶ Bankruptcy, Economics of
- ▶ Inside and Outside Money
- ▶ Money and General Equilibrium
- ▶ Network Goods (Theory)

Bibliography

Ashtor, E. 1972. Banking instruments between the Muslim East and Christian West. *Journal of European Economic History* 1: 553–573.

Baxter, W. 1983. Bank exchange of transactional paper: Legal and economic perspectives. *Journal of Law and Economics* 26: 541–588.

Bliss, R. 2003. Bankruptcy law and large complex financial organizations: A primer. *Federal Reserve Bank of Chicago Economic Perspectives* 27(1): 48–58.

Cavalcanti, R., and N. Wallace. 1999. A model of private banknote issue. *Review of Economic Dynamics* 2: 104–136.

Committee on Payment and Settlement Systems. 2005. *Statistics on payment and settlement systems: Figures for 2003*. Basel: Bank for International Settlements.

Evans, D., and R. Schmalensee. 1999. *Paying with plastic: The digital revolution in buying and borrowing*. Cambridge, MA: MIT Press.

Evans, D., and R. Schmalensee 2005. *The economics of interchange fees and their regulation: An overview*. Working Paper No. 18181. Cambridge, MA: Sloan School of Management, MIT.

Freeman, S. 1996. The payments system, liquidity, and rediscounting. *American Economic Review* 86: 1126–1138.

Green, E. 1999. We need to think straight about electronic payments. *Journal of Money, Credit, and Banking* 31: 668–670.

Kahn, C., and W. Roberds 2001. *Transferability, finality, and debt settlement*. Working paper. Federal Reserve Bank of Atlanta.

Kahn, C., and W. Roberds 2003. *Payments settlement under limited enforcement: Public versus private systems*. Working paper. Federal Reserve Bank of Atlanta.

Kahn, C., J. McAndrews, and W. Roberds. 2003. Settlement risk under net and gross settlement. *Journal of Money, Credit, and Banking* 47: 299–319.

Kiyotaki, N., and J. Moore 1997. *Credit chains. Mimeo*. London School of Economics.

Kiyotaki, N. and Moore, J. 2000. *Inside money and liquidity. Mimeo*. London School of Economics.

Kohn, M. 1999. *The capital market before 1600*. Working Paper No. 99-06. Department of Economics, Dartmouth College.

Kohn, M. 2001. *Payments and the development of finance in pre-industrial Europe*. Working Paper No. 01-05. Department of Economics, Dartmouth College.

Rochet, J.C. and Tirole, J. 2004. *Two-sided markets: An overview*. Working Paper No. 275. IDEI, University of Toulouse.

Weinberg, J. 1997. The organization of private payment networks. *Federal Reserve Bank of Richmond Economic Quarterly* 83(2): 25–43.

Pay-off Period

D. M. Nuti

The *pay-off period* of an investment project is the number of years over which the project pays for itself from the time of completion, i.e. the sum of undiscounted after-tax gross profits over the period are equal to total investment outlays. There is evidence that enterprises investing in plant and equipment, mostly in industry, require for a project to be undertaken that its pay-off period should be no longer than a *standard* period which is customary in the given sector of operation, ranging from under two to five years. If mutually exclusive projects are available, for instance if there are alternative techniques of production available for creating otherwise identical productive capacity, *ceteris paribus* the pay-off period of investment is not minimized but is brought closest to the standard pay-off period of the sector involved, which is not to be exceeded. It is an *average satisfying* condition, not a marginal condition for optimization. Its satisfaction for any given investment and current costs associated with it can be ensured by a corresponding appropriate mark-up on current costs in output pricing.

Early evidence of this kind of investment behaviour was documented by Henderson (1938), Meade and Andrews (1938), Brockie and Grey (1956). In general this criterion of investment choice tends to be discussed in business textbooks rather than in economic theory; a notable exception is Kaldor and Mirrlees (1962), where this investment criterion plays a major role in determining the economy's growth path.

The pay-off period criterion would appear to be at odds with discounted cash flows methods, whereby the present value of investment, i.e. the cumulative sum of discounted net cash flows at start, should be maximized, being tantamount to the increase in net wealth deriving specifically from undertaking the investment. Discounting methods lead to projects being undertaken if their present value is positive or – which mostly but not always gives the same result but not necessarily the same ranking of projects – if the discount rate expressing the opportunity cost of finance to the investor is lower than the internal rate of return on investment (i.e. the rate which, if it exists and is unique, would make the present value of investment equal to zero at the point of starting the investment; we neglect here the questions of multiple internal rates of return, variable cost of finance, optional investment lifetime and other complications of discount methods). The divergence between pay-off and discounting criteria is not, however, as great as it may seem.

First, usually the pay-off criterion is applied only to investments which satisfy the additional criterion of earning a minimum rate of profit over the investment lifetime higher than a target rate related to the opportunity cost of finance to the investor. However, the pay-off criterion is likely to be a stricter test than the minimum profit rate; Kaldor and Mirrlees (1962) assume that whenever the first is satisfied, the second is automatically satisfied.

Second, discounting procedures apply primarily to a world of certainty: to handle uncertainty those procedures require either adding a percentage risk factor to the interest rate, which unduly amplifies the weight attached to the riskiness of future distant events; or the comparison of 'certainty equivalent' present values, i.e. a purely subjective trade-off between mean present value and its standard deviation (or other measure of dispersion). The pay-off criterion is a rough and ready way of handling uncertainty, in particular about potential competition generated by new technology. The best explanation of the pay-off criterion, owed to Kaldor, is that the standard period is a parameter chosen by firms on the basis of experience in such a way as to meet the uncertainty due to obsolescence in different sectors and the time required by the introduction of new techniques. An entrepreneur undertaking a new investment is subject to the risk that technical advance in his field will make his investment obsolete cutting the flow of his profits. Cheaper substitutes for his product may be introduced, or he may be unable to take advantage of subsequent more efficient methods of making his product once he has committed his investment funds to a specific technical form embodying the best current practice. But it takes a certain number of years to develop a new process up to the point of industrial application on a large scale, and since he knows that for the time being there are no such new processes, he is prepared to invest in projects from which he expects to get his own money back within that number of years, whereas he is reluctant to risk investments which would normally pay for themselves over a longer period.

In the Soviet Union and other East European countries in the mid- and late-1950s official instructions were issued for the selection of investment projects which are formally similar to the pay-off criterion but are radically different in both theoretical meaning and empirical justification. Basically, if two alternative ways of producing the same new capacity were available, the more investment intensive project would be selected only if its current operating costs were so much lower that the *additional* investment expenditure could be recouped through current costs savings within a given number of years T fixed as 'standard' by central planners. Another way of putting this is that the more investment intensive project would be selected only if the sum of its operating costs and $1/T$ of its investment outlays were lower than for the other project; this leads naturally to the generalization of the principle, for any given new capacity to be generated, as the minimization of the sum of operating costs and $1/T$ of investment outlays. Thus investors obeying this rule effectively behaved *as if* they were subject to a capital charge equal to $1/T$ in spite of the fact that investment funds were obtained free of charge from the state budget (only subject to reimbursement of straight line depreciation). Given a uniform lifetime n of the investment plants a standard period T defines a shadow interest rate r implicit in the relationship

$$T = \frac{(1+r)^t - 1}{r \cdot (1+r)^t}$$

The standard period T in English-language literature is usually labelled 'recoupment period' (*rok okupoemosti* in Russian, *doba navratnosti* in Czech, *czas zwrotu* in Polish) and in spite of the non-existent nuance of meaning between 'recoupment' and 'pay-off' the maintenance of this convention to distinguish between the two concepts is essential in view of their radical differences.

The Soviet-type standard recoupment criterion is a marginal, not an average condition and is part of an optimization (i.e. cost minimization) procedure in an economy where otherwise there are no capital charges. It does not regulate whether or not the new capacity should be generated but it only regulates its technical form; thus it is insensitive to the relative price of output and inputs. For a single investment project a recoupment period cannot be defined and is not necessary once planners have decided to create that new capacity. The pay-off criterion of the capitalist firm on the contrary is an average not marginal condition and can only be regarded as an approximate rule of thumb in lieu of optimization. The pay-off period of a single investment can be defined and is certainly sensitive to the relative price of inputs and outputs. It regulates both whether new capacity is established and its technical form. The two criteria have in common the comparison of items of capital stocks and undiscounted flows and therefore also the common time unit but should not be confused (as they often have been in comparative literature).

The recoupment period criterion in the Soviet Union was officially codified in official investment regulations in 1960 and 1962, but had already been informally used in the infraministerial distribution of investment funds since the mid-1930s (for a discussion of early practices see Hunter 1949) as a way of ensuring efficiency in the absence of interest charges on investment (on early Soviet discussions on interest and capital charges see Grossman 1953). Similar criteria were introduced in Czechoslovakia (1961), Poland (1963) and Hungary (1963). Soviet and Czech rules added straight line depreciation percentages to $1/T$ in comparing investment alternatives; Polish and Hungarian rules did not. Soviet and Czech rules had different 'standard' recoupment periods in different industries, ranging from 4 to 10 years. The criterion rules on the technical form of investment and not on whether new capacity should be created at all, thus diversified recoupment periods by sectors cannot be regarded as an expression of central planner's sectoral priorities and have no justification. It is probable that in Soviet practice in each industry a standard period emerged to rationalize technical choice by engineers and projectdesigning organizations, but that softer financial constraints led to longer standard recoupment periods in priority industries – a practice later codified in both the Soviet Union and Czechoslovakia without consideration of the reasons for it. Polish procedures (1962) are the most elaborate and contain allowances for the freezing of investment resources during gestation, allowances for the impact of durability on both costs (given embodied technical progress) and output as well as a justification for the determination of the standard recoupment period (6 years throughout the economy on new investment) which was related to the recoupment period of labour-saving investment in modernization, thought to be recoverable in five years on a large scale (the difference between 5 and 6 years being accounted by the shorter lived nature of benefits deriving from investment in modernization). The original elements of the Polish investment rules were due to the work of Michal Kalecki and Mieczyslaw Rakowski (1959) which they embodied almost verbatim. In other East European countries similar rules were apparently used more or less officially (Katchaturov 1962; for a general discussion of these criteria see Dobb 1951; Zauberman 1955, 1962; Nuti 1971; for a criticism of the Kalecki–Rakowski approach see also Nuti 1986).

Harcourt (1968, 1969) has compared the impact of alternative investment rules such as discounted cash flow methods and recoupment periods. In the late 1960s and early 1970s new rules were issued in Czechoslovakia and Poland (see Nuti 1970, 1971) approaching more conventional Western type discounted cash flow methods – including emphasis on international prices as in standard OECD and UNIDO cost/benefit analysis, while the 1972 Soviet investment regulations maintained the earlier approach with small modifications. The 'recoupment period' approach – whatever the official permanence of regulations – has been made redundant by the appearance of interest-bearing investment credit (in Hungary and Poland on a large scale). Economic crisis in Eastern Europe at the turn of the 1980s has introduced emphasis on quickyielding investment, with versions of the capitalist pay-off criterion (especially for outlays and revenues respectively incurred and accrued in hard currencies) used to allocate investment funds among competing firms. This however is an indication of underpriced hard currency and external imbalance; it does not have the same rationale given above (i.e. technological uncertainty) for the capitalist pay-off criterion and it is part of an optimization attempt (maximize the flow of net export receipts) instead of being an average condition.

See Also

- ▶ Internal Rate of Return
- ▶ Investment Decision Criteria

Bibliography

Akademiya Nauk SSSR. 1960. *Tipovaya Metodika Opredeleniya Ekonomicheskoi Effektivnosti Kapital'nykh Vlozhenii i Novoi Tekhnicki v Narodnom Khozyaistve SSSR* (Standard methodology for determining the economic effectiveness of capital investment and new technology). Moscow: Gosplanizdat.

Brockie, M.D., and A.L. Grey. 1956. The marginal efficiency of capital and investment programming. *Economic Journal* 66: 662–675.

Dobb, M.H. 1951. A note on the discussion of the problem of choice between alternative investment projects. *Soviet Studies* 2(3): 289–295.

Dunlop, J.T., and N. Fedorenko (eds.). 1969. *Planning and markets: Modern trends in various economic systems*. New York: McGraw-Hill.

Gosplan i Akademiya Nauk SSSR. 1962. *Metodika Opredeleniya Ekonomicheskoi Effektivnosti Vnedreniya Novoi Tekhniki, Mekhanizatsii i Avtomatizatsii Proisvodstvennykh Protsessov v Promyshlennosti* (Methodology for determining the economic effectiveness of the introduction of new technology, mechanization and automation of productive processes in industry). Izdatelstvo Akademii Nauk SSSR: Moscow.

Grossman, G. 1953. Scarce capital and Soviet doctrine. *Quarterly Journal of Economics* 67(3): 311–343.

Harcourt, G.C. 1968. Investment decision criteria, investment incentives and the choice of techniques. *Economic Journal* 78: 77–95.

Harcourt, G.C. 1969. Investment decision criteria, capital intensity and the choice of techniques. In Dunlop and Fedorenko (1969).

Henderson, H.D. 1938. The significance of the rate of interest. *Oxford Economic Papers* 1: 1–13.

Kaldor, N., and J. Mirrlees. 1962. A new model of economic growth. *Review of Economic Studies* 29: 174–192.

Kalecki, M., and M. Rakowski. 1959. Uogolnienie wzoru ekonomicznej efektywnosci inwestycji (A generalized formula of investment effectiveness). *Gospodarka Planowa*, no. 11; English translation in Nove and Zauberman (1964).

Katchaturov, T.S., ed. 1962. *Voprosy ekonomicheskoi effektivnosti kapitalnovlozhenii* (Problems of the economic effectiveness of investment). Moscow.

Komisia Planowania Przy Radzie Ministrow. 1962. *Instrukcja Ogolna w sprawie metodyki badan ekonomicznej efektywnosci inwestycji* (General instructions on the methodology for determining the economic effectiveness of investment). Warsaw.

Meade, J.E., and P.W.S. Andrews. 1938. Summary of replies to questions on effects of interest rates. *Oxford Economic Papers* 1: 14–31.

Nove, A., and A. Zauberman (eds.). 1964. *Studies in the theory of reproduction and prices*. Warsaw: PWN.

Nuti, D.M. 1970. Investment reforms in Czechoslovakia. *Soviet Studies* 21(3): 360–370.

Nuti, D.M. 1971. The evolution of Polish investment planning. *Jahrbuch der Wirtschaft Osteuropas*, Vol. 3, Munich/Vienna.

Nuti, D.M. 1986. Michal Kalecki's contributions to the theory and practice of socialist planning. *Cambridge Journal of Economics* 10(4): 333–353.

Orszagos Tervhivatal-Penzugyminiszterium-Epitesugyi Miniszterium. 1963. *Beruhazasi Kodex* (Investment Code). Budapest.

Statni Planovaci Komise. 1961. *Smernice o urcovani efektivnosti investic a nove techniky v narodnim hospodarstvi CSSSR* (Directives for the assessment of the effectiveness of investment and new techniques in the national economy of the CSSE). 10 April, Prague.

Zauberman, A. 1955. A note on Soviet capital controversy. *Quarterly Journal of Economics* 69(3): 445–451.

Zauberman, A. 1962. The Soviet and Polish quest for a criterion of investment efficiency. *Economica* 29: 234–254.

Payroll Taxes

Daniel S. Hamermesh

Taxes paid by employers based on the number of employees and the wages paid. These taxes are constructed by applying a *tax rate* to the per-period wage rate paid to the employee. The rate may differ among employees or across firms. In a few tax structures there is a *ceiling* on the total payroll tax bill assessed against the earnings of each employee.

Although historically not an important source of revenue for national governments, they became a major and increasingly important source of revenue for national governments after World War II. By 1980 payroll taxes accounted for 15 per cent of the tax revenue of the Federal government in the United States, 13 per cent in Great Britain, and 39 per cent in Sweden. Payroll taxes have several purposes. In many cases a payroll tax is explicitly designed to finance programmes viewed as benefiting workers. For example, payments to retirees under the United States Social Security Act are financed by a payroll tax (as well as by an earnings tax on workers). The tax may also be an instrument of economic policy, either *experience rated* – varying in amount with each firm's past record of generating benefits financed by the tax – or raised or lowered to affect employment over the cycle, across areas or among demographic groups.

The major economic issue of interest in the payroll tax is its *incidence* (see Musgrave and Musgrave 1980, for a discussion of incidence in a more general context). Stated most broadly, the question economists must answer is: what is the effect on the time path of factor and product prices and quantities of an increase in the tax assessed on the firm against the payroll of the workers it employs?

The most widely discussed issue under this general question is whether labour bears the payroll tax. (This is generally taken to mean whether net per-worker earnings after the payroll tax is assessed are equal to their pre-tax level minus the tax.) If the supply of labour to the market is completely inelastic, labour bears the entire tax; tax-induced shifts in labour demand merely move the demand curve along the vertical labour-supply curve. If the supply elasticity is non-zero, the elasticity of demand for labour will also affect the outcome. On *a priori* grounds the answer thus depends on one's beliefs about the empirical magnitudes of labour supply elasticities. Despite the centrality of labour-supply behaviour in resolving this question, most empirical research has focused on attempts to estimate directly the effect of differences in payroll tax rates on wage rates across subunits and over time. Thus one leading empirical study (Brittain 1972) estimates a CES production function across countries, modified to incorporate differences in payroll tax rates and their incidence, and finds full shifting onto labour. Other empirical studies, using data on individuals whose employers pay different taxes on their earnings because of ceilings on the tax (Hamermesh 1979), or using aggregate time-series data to examine how average wage changes respond to payroll tax changes (Beach and Balfour 1983), reach sharply differing conclusions. The conflicting evidence forces one back onto extraneous estimates of labour supply elasticities in order to reach conclusions about the burden of this tax. Since the best estimates of these suggest they are small but positive, and since the aggregate demand elasticity for labour is below unity, we may infer that labour bears most, but not all of the tax in the form of net wages only slightly greater than the pre-tax wage less the tax.

A second major issue is the effect of increased payroll taxes on the price level. In the popular press, and among some Keynesian economists, the increases are viewed as 'passed on' in the form of higher product prices. These views ignore the effect of monetary policy and spending policies on aggregate outcomes; they also implicitly assume that the aggregate demand is price inelastic. Payroll tax increases represent a negative supply shock that can be easily analysed in the standard aggregate demand/aggregate supply framework. As such, they produce a temporary decline in output and rise in price inflation, both of which are eventually removed as price expectations adjust. As with other negative supply shocks, the temporary decline in output can be mitigated, at the cost of more rapid inflation, by expansionary spending or monetary policy. Payroll tax increases are 'passed on' in the form of higher product prices, unless net wages fall by an amount equal to the tax increase. But the effect on prices is temporary unless the government accommodates the increase by stimulating aggregate demand. The arguments on the effects of payroll taxes on the macroeconomy hold in reverse for payroll subsidies.

A third issue under the general question of incidence is the effect of the tax on the distribution of net incomes. This problem is frequently dealt with in popular discussions of payroll taxes that embody ceilings, as such taxes are superficially regressive. Such discussions implicitly assume that labour bears the tax, and also assume the burden to be proportionate to the tax each particular worker generates. Perhaps even more important, they ignore the common link, either explicit or implicit, between the tax and the distribution of benefits of the programmes it finances. These are best viewed as a package; and there is no satisfactory evidence on the effect of the tax/benefit package on net incomes.

Most analyses of the incidence of payroll taxes are static and, like the discussion here, implicitly long-run. Even if the tax is eventually borne mostly by workers because the long-run supply of labour to the market is relatively inelastic, and even if there are no net effects on income distribution, labour-market dynamics can produce quite different short-run outcomes (Hamermesh 1980). While labour demand adjusts fairly rapidly to its long-run equilibrium, labour supply adjusts slowly because of lags in perception and in training. These lags are sufficient to make the incidence of payroll tax increases rest less heavily on workers for several years after they are instituted, and to give employers more incentive to attempt to raise prices (thus producing a larger negative supply shock).

Leaving the general question of incidence, we also know that the existence of different marginal payroll tax rates assessed on workers with different earnings rates and of ceilings on taxable earnings means that changes in payroll taxes will affect the mix of labour inputs. Thus, for example, an increase in the rate of a tax with a low ceiling effectively increases the fixed costs of employing workers, but does not raise the cost of adding another hour to the work week. It thus induces employers to substitute away from workers and toward hours along an isoquant. Similarly, a higher ceiling with an unchanged rate raises the relative cost of employing more skilled workers and induces substitution away from them, especially given the strong evidence that they and less skilled workers are substitutes in production. These and other examples indicate that payroll tax policy can be used as a tool of labour-market policy; and policy analysts have increasingly recognized that the impact of changes in the structure of payroll taxes on employment must be considered.

Experience-rated payroll taxes are designed to reduce the incidence of the activity that generates the payments the taxes finance. Periods of unemployment and workplace injuries are the best-known examples of these activities. Increasing the degree of experience rating in a payroll tax that finances these benefits will tend to reduce the incidence of the activity. A mass of empirical work is fairly conclusive on the validity of this theoretical proposition, though the range of estimates is so wide that one cannot infer the magnitude of the reductions in injuries and unemployment that could be induced by better experience rating of taxes. Instituting a system of benefits financed by a payroll tax that is not experience rated may actually *increase* the incidence of the loss. To the extent that workers derive some utility from the consumption of leisure financed by unemployment benefits, they will supply their labour more cheaply to the firms that offer such leisure. The combination of lower wages and the same payroll tax bill leads such firms, those that generate substantial unemployment, to expand, thereby increasing the total variability of employment.

See Also

► Social security

Bibliography

Beach, C.M., and F.S. Balfour. 1983. Estimated payroll tax incidence and aggregate demand for labour in the United Kingdom. *Economica* 50(197): 35–48.

Brittain, J. 1972. *The payroll tax for social security*. Washington, DC: Brookings Institution.

Hamermesh, D. 1979. New estimates of the incidence of payroll tax. *Southern Economic Journal* 45(4): 1208–1219.

Hamermesh, D. 1980. Factor market dynamics and the incidence of taxes and subsidies. *Quarterly Journal of Economics* 95(4): 751–764.

Musgrave, R., and P. Musgrave. 1980. *Public finance in theory and practice*. New York: McGraw-Hill.

Peacock, Alan T. (1922–2014)

Martin Ricketts

Abstract

The central concern of Alan Peacock's work was in public finance. He defended the traditions of classical liberal political economy against the claims of 'The New Welfare Economics' to represent the only theoretically rigorous approach to public policy. His contribution was not focused on either positive or normative economics, but on what John Neville Keynes (1891. *The scope and method of political economy*, 1st ed. London: Macmillan) called the 'art of economics' – the translation of normative principles into policy action given the constraints presented by practically feasible institutions governing both private and public choice. His policy interests were wide and ranged from the big questions of the size and scope of government to more specific issues such as the structure of the social security system, the finance of education, the public support of the arts generally and the finance of the BBC in particular.

Keywords

Arts funding; Bargaining with regulators; BBC licence fee; Bureaucracy; Classical liberalism; Devolution; Displacement effect; Education vouchers; Growth of public expenditure; Heritage; Italian tradition in public finance; Ordoliberalism; Public choice; Rowley, C.; Welfare economics; Welfare state; Wiseman, J.

JEL Classifications

B12; D6; D72; D73; D78; H1; H2; H5; H7; I22; I28; I38; Z11; Z18

Born in Ryton on Tyne, Alan Peacock was educated in Dundee after his father moved from Armstrong College in Newcastle to become Professor of Zoology at the University of St Andrews. Alexander Peacock was an entomologist who had studied mosquitoes in West Africa and the lice that were the cause of trench fever while he served as a soldier on the Western Front during the First World War. Alan Peacock grew up, therefore, in an environment that heavily emphasised the social responsibilities attached to receiving a higher education and, as he wrote in a 'quasi-autobiographical' book *Anxious to do Good* (2010, p. 10), his father 'was a particularly hard act to follow so far as "doing good" is concerned'. His parents' desire for reconciliation with the Germans after the horrors of the First World War led to holidays in Germany and to his studying German at school – a fact that played a significant role in Peacock's later career.

Peacock's studies at the University of St Andrews (1939–42) were interrupted by the Second World War. His knowledge of German led to the role of a sea-going Intelligence Officer of the 'Y' service in the Royal Navy, decoding enemy communications, an account of which appeared as *The Enigmatic Sailor* (2003). In this capacity he took part in Operation Tunnel, in which his ship, HMS *Limbourne*, was torpedoed and sunk along with HMS *Charybdis*. The likelihood of available information being overlooked or misinterpreted had a significant effect on the young Peacock (the captain of *Charybdis,* for example, had not fully appreciated the source of information from *Limbourne* and mistook 'Y' information for the less technically specific 'my' information [Peacock (2003, p. 41)]. Peacock went on to serve on operations protecting the Arctic Convoys taking assistance to Russia via Murmansk and was awarded the Distinguished Service Cross in 1945.

After the war he completed his degree in Economics and Political Science at St Andrews and, following a brief spell as a lecturer there, moved to the London School of Economics in 1948 where he was influenced by Lionel Robbins and Friedrich Hayek. In later years he was to become Reader in Public Finance at the LSE (1951) and to hold Professorships at Edinburgh (1956–62), York (1962–78) and Buckingham (1978–84). He remained Professor Emeritus at Buckingham until his death. He also played a large part in establishing the David Hume Institute at Edinburgh.

The Growth of Government

Peacock's early work was on National Insurance and the analysis of the Beveridge Report (1942) that formed the basis for the growth of the post-war Welfare State in the UK. Beveridge was a Liberal academic at the London School of Economics from a tradition that emphasised personal independence and responsibility, but his reforms were implemented by a Labour administration in difficult economic circumstances and in a world that had become increasingly used to high levels of state intervention and planning. Coming from Scottish Liberal political circles, Peacock was concerned with the problem of reconciling the aims of social security provision with the maintenance of a liberal economic order in the classical sense.

In the post-war environment of the UK, critical analysis of social policy did not rank highly in academic circles. Peacock (2010, p. 45) records that in the period 1925–1955, of one thousand articles in the *Economic Journal*, only ten concerned the social services. He therefore began researching the institutional background and financing of pensions, unemployment insurance, 'national assistance' (as transfers to those in poverty were then termed), and provision for sickness and disability. Beveridge had recommended an 'insurance' based system of social support with premiums payable by all (if necessary supported by the taxpayer). By the 1950s it was clear, however, that the state, whether controlled by Labour or Conservative governments, was to play a dominant role on the supply side and that the system as a whole was becoming highly centralised and collectivised. It was also becoming clear that the National Insurance Scheme was not simply about creating 'insurance' markets that might not evolve without state encouragement, but was part of a broader policy towards the redistribution of income.

Criticism of both these trends provides the two main leitmotifs of Peacock's professional writing. His Scottish Liberalism as well as his classical liberalism led to profound suspicion of state monopoly provision as well as a preference for redistributional instruments that gave the recipients a degree of freedom of choice and the chance to avoid dependency. Much of his early work involved mastering the national accounts (still a relatively new framework for the collection and organisation of economic data at the time) and considering how these would be affected over time by demographic trends, changes in the rules of entitlement to benefits and trends in economic growth. This resulted in an early paper on the National Insurance Funds (1949) in which he proposed integrating National Insurance with the accounts of the public sector and giving up 'the pretence that the scheme is based on insurance principles'. This was followed by a more extensive study questioning the National Insurance system (1952), an article projecting the likely strain on the state's budget up to the 1980s of existing policies (1954a, with Frank Paish) and a book specifically on the techniques of national income accounting (1954b, with Harold Edey).

During this early period Peacock had also been involved as a Liberal party advisor. He was a member of the 'Unservile State Group' of academics and others sympathetic to the Liberal Party, which was set up in 1953 and produced a report entitled *The Unservile State: Essays in Liberty and Welfare* in 1957. He also provided research assistance for a report (1950) on the *Reform of Income Tax and Social Security Payments*. This was based around the work of Lady Juliet Rhys Williams and proposed a 'negative income tax' (later associated with Milton and Rose Friedman (1962)) by which minimum levels of income might be guaranteed to families through the normal operation of the fiscal system. An allowance would be paid to families (whether in or out of work) financed out of the taxation of income in excess of this level. It drew stark attention to the implied trade-offs between higher minimum standards and the possible adverse effects on work incentives and growth of the tax rates required to achieve them. Peacock was involved in providing estimates of the tax rates required to finance the scheme – though estimates of the incentive consequences were highly conjectural. The Liberal Party report proved to be at odds with the temper of the times and was considered but rejected by the Royal Commission on the Taxation of Profits and Income that received evidence in 1951. The old radical and individualist traditions going back to John Stuart Mill were increasingly supplanted by a more interventionist and paternalist philosophy, both within the Liberal Party and more generally.

Peacock's interest in social policy and involvement in Liberal politics had brought him to consider a major component of the growth of public expenditure. As an academic public finance economist he was also interested in placing these developments in redistributional finance in the context of normative and positive theories of government growth more generally. With Richard Musgrave (1958) he edited a collection of *Classics in the Theory of Public Finance*. This included contributions from Adolf Wagner (1883), Knut Wicksell (1896) (translated by the later Nobel prize-winning economist James Buchanan), Erik Lindahl (1919) and other articles translated from German, Italian and French. Samuelson (1954) had clarified the technical conditions for the efficient provision of 'public goods', but the older literature, with its emphasis on consensus (Wicksell) or bargaining over cost shares (Lindahl), draws greater attention to the institutional problems of collective choice and the possibility that actual results will depart substantially from the ideal.

Peacock wrote with Jack Wiseman (1961) a study of the *Growth of Public Expenditure in the United Kingdom* which charts the historical trends and considers the various possible explanations for them. In particular, this study is celebrated for its critique of Wagner's 'law' of increasing state activity (pp. 16–24) and the investigation of two hypotheses – the 'displacement effect' and the 'concentration process' (pp. 24–30). Peacock and Wiseman were critical of some of Wagner's arguments for a rising share of public in total expenditure – such as the inherent superiority of public over private corporations in producing economic stability. But they advanced alternative explanations of their own for particular periods of public sector growth in the UK. Public expenditure was 'displaced' upwards after periods of social stress, such as wars or severe social disturbance. Higher levels of taxation were tolerated under extreme conditions 'and this acceptance remains when the disturbance itself has disappeared' (p. 27). The concentration process referred to a tendency for larger and more centralised delivery mechanisms to evolve partly in response to economies of scale, but also to political demands for uniformity of standards.

These two hypotheses seemed descriptively to fit the story of 20th century developments in the UK. But, as Peacock and Wiseman themselves emphasised, they did not constitute a fully developed theory of government growth. The main influence on Peacock's work during and after the 1960s was the advance of public choice theory associated with Downs (1957), Buchanan and Tullock (1965), Breton (1974), Olson (1974) and Niskanen (1971). These offered a much more systematic analysis of collective choice by looking at the individual incentives faced by economic agents within a range of constitutional or organisational settings.

Two factors were paramount in Peacock's later writing about the growth of government. The first was the tendency within democracies for particular interests and pressure groups to divert resources in their favour through the regulatory as well as the fiscal system. On the demand side, this was partly the result of the ability of relatively poor voters in a democracy to load the tax costs of public activities onto richer voters through progressive tax systems. But it was also related to the ability of smaller cohesive interest groups, facing relatively low costs of collective action, to exert pressure on politicians and legislators in representative democracies. The second important factor concerned the 'supply' side of the political market. Legislators and bureaucrats were themselves operating within an institutional and organisational context that would be expected to lead to expanding public spending.

Peacock's work in Public Choice provides an interpretative commentary rather than theoretical novelty. As a classical liberal he was always interested in the historical origins of ideas about the role and growth of the state. He wrote a set of lectures (1992a) on *Public Choice in Historical Perspective* in which he pointed to de Tocqueville's (1965) work (also in Peacock (1978a)) as a prologue to much modern theory (pp. 42–53). In the same lectures he returned to the Italian tradition in public finance, which he much admired, and in particular the contribution of Amilcare Puviani (1903) on fiscal illusion (pp. 96–102).

Although he would occasionally specify a theoretical model in formal mathematical terms, his efforts were usually directed at discussing the problems of interpretation, the difficulties of defining and measuring the arguments of the utility or production functions used, the possibility that important elements of the problem had been omitted and so forth. The Economics of Bureaucracy, for example, might find it convenient to summarise the objective of senior administrators as maximising their budgets. Peacock (1978b) accepted that this might be a useful first approximation (although still probably culturally specific), but matters such as reputation with colleagues, professional pride, the desire for a quiet life, pure laziness, personal ambition or an idealistic wish to serve the public needed to be borne in mind when applying the theory in particular cases.

It would be a mistake to regard Peacock's commentary on Public Choice as representing an attack on formal theory by the easy paths of questioning assumptions, complaining about lack of completeness or lack of descriptive realism. He admired many of the theorists, quoted their work extensively and was well aware of methodological defences to quite abstract and 'unrealistic' theory. His contribution, however, was not in positive theory itself, but in using the theory to cast light on the whole process of policy formulation and execution. This 'art of economics' requires a more classical approach to the subject. Peacock's economic agents whether business people, voters, politicians, bureaucrats, lobbyists or indeed academics were those of Adam Smith – broadly self-interested, but hard to sum up in a simple utility function – and 'the situation in which they are placed', as Smith put it, was more complex than could be summarised in a set of known resource constraints.

Lack of information or the possibility that it might be systematically distorted also played a large part in Peacock's view of policy leading to an awareness, more normally associated with 'Austrian' economists, of the unintended consequences of intervention. He was also wary of theoretical approaches that viewed people as 'passive adjusters' to policy instruments instead of active bargainers. This can be seen in the titles of articles and books – *The Regulation Game* (1984), *The Heritage Game* (2008), 'Bargaining and the regulatory system' (1986) – and in papers (1981) emphasising that taxpayers can respond not merely by avoidance or evasion, but with various forms of political action.

The Critique of Welfare Economics

Peacock's classical liberal approach to public policy had a direct bearing on his suspicion of Welfare Economics. By the 1950s the New Welfare Economics deriving from Vilfredo Pareto, based upon the inadmissibility of interpersonal comparisons of utility and the concept of 'Pareto efficiency', had become the established paradigm in normative theory. It had been saved from its seemingly highly restrictive applicability (most policy changes would make some people worse off and others better off, thus requiring the forbidden interpersonal comparisons of utility in any final assessment) by the development of the Hicks (1939)–Kaldor (1939) compensation tests. Where the gainers from a change from one social state to another could compensate the losers there was a 'potential Pareto improvement' in social welfare, and it increasingly became accepted (in practical policy discussion if not in purely philosophical discussion) that changes of this nature were desirable (even if compensation were not actually paid). In essence, the pursuit of 'economic efficiency' came to be treated as a dominant aim of public policy, while the ability to characterise these Pareto efficient states in mathematical terms lent a spurious 'scientific' credibility to the resulting policy proposals.

The value judgements underlying Paretian welfare economics were widely considered very weak and indeed perfectly consistent with liberal principles. However, the pursuit of 'efficiency' through public policy was not always consistent with the tenets of classical liberalism conceived of as the defence of negative freedom. Much of Peacock's writing is based on this essential point. In Peacock and Rowley (1975) the theme was set out at length. 'It is an illusion to infer that Paretian welfare judgements lead to non-authoritarian solutions in matters of economic policy simply because they are derived from individual preferences' (p. 2). The whole book is an attempt to re-establish a 'classical political economy with modern trappings' (p. 3) against the claims of Paretian welfare economics that had established a 'dangerous hegemony in the economic theory of public policy'.

Ultimately the critique is founded on a rejection of 'welfarism' – the idea that social welfare should be derived entirely from the self-assessed utility of individuals. In Peacock (1992a, p. 104), for example, he explicitly remarks that 'it is amazing to me how often it is taken for granted that if by some miraculous device, all preferences for social goods could be known . . . an imposed solution is then considered justified'. The actual taking part in a process of political decision-making is important – just as the exercise of judgement in matters of private exchange is important in developing people with qualities that permit the exercise of personal responsibility. The use of non-utility information to assess policy involves value judgements quite distinct from the idea that 'individuals are the best judges of their own welfare'. Outcomes alone (and their consequences for the utility of individuals) are not all that count from a liberal perspective or indeed from the perspective of other radical and 'New Left' critiques of static welfare theory (see Peacock and Rowley 1975, pp. 69–76).Sen's (1970) celebrated proof of the impossibility of a Paretian liberal, while clearly drawing attention to potential conflicts between formal social choice theory and liberalism, is nevertheless criticised by Peacock and Rowley (pp. 80–3) for advancing an inadequate condition of 'minimal liberalism'.

Although Peacock saw the classical liberal approach to both private and public choice as supporting the claims of negative liberty, he was not a fully fledged 'process liberal', such as Ludwig von Mises. Mises and others of the 'Austrian School' conceived economic life as entirely about change and dynamic market processes. These processes would never come to rest and the time path could not be determined in advance. Normative assessment could therefore not be in terms of the achievement of particular end states, but had to rely on the compatibility of the process itself with desirable features – such as the maintenance of an open society and individual liberty.

Peacock had greater affinity with the German neo-liberals, whom he categorised as 'end state' liberals (Peacock and Willgerodt 1989b, p. 3). His knowledge of German and his acquaintance after the war with Herbert Giersch at the LSE introduced him to the work of Walter Eucken and other 'ordo-liberals'. Ludwig Erhard, one of the main architects of West German post-war reform, was associated with the neo-liberals and is widely credited with the implementation of the *Soziale Marktwirtschaft* (a term coined by Alfred Müller-Armack). Peacock and Willgerodt (1989a) edited a set of translations of some of the major contributors to this German neo-liberal tradition including Walter Eucken, Franz Böhm, Wilhelm Röpke, Alfred Müller-Armack and others.

The distinguishing feature of the ordo-liberals compared with the more libertarian followers of Mises and Hayek was their conviction that the state had a crucial role to play in ensuring that various undesirable outcomes were avoided. A central task, for example, was to ensure that market processes did not lead to excessive accumulations of private monopoly power. Such power was dangerous because it could be used to establish political privilege and undermine the competitive liberal order. Mises and others might argue that no position was immune from new entry and technical change in the very long run, but this would not be true if the state itself began to fall under the influence of powerful private interests. Similarly, intervention to try to place some limits on income inequality was accepted as a necessary part of maintaining political support for an essentially liberal economic system. This went further than measures to relieve abject destitution, but not as far as a comprehensive 'social welfare state' – which itself would be a threat to liberalism (Peacock and Willgerodt (1989b, p. 12). Maintaining a reliable and non-inflationary currency was also a matter of central importance to the neo-liberals, as the disastrous consequences for liberalism of the Weimar inflation were still a matter of recent experience.

Where, or within what limits, in the interventionist spectrum the liberal order might find a sustainable location was a matter for continuing investigation and argument. But the desire, for example, to maintain competition by preventing mergers above a certain size or by forbidding particular restrictive agreements (even if these were entirely voluntary in nature) implies that certain types of social 'outcomes' matter. In this sense classical liberals such as Adam Smith and John Stuart Mill were 'end state' liberals even if their ultimate interest was to defend a system of 'natural liberty' that gave maximum scope for market processes and the exercise of individual choice.

Peacock fitted into this 'classical' historical tradition. It was a tradition deeply suspicious of Welfare Economics. His support for competition, for example, did not derive from the First Theorem of Welfare Economics that a perfectly competitive equilibrium (in the absence of externalities) was Pareto efficient. Nor did he think that it made any sense for policy to aim at approximating perfectly competitive conditions – which would have the effect of expunging conscious rivalry from the economic system. The case for competition in classical liberalism was based on a suspicion of concentrated private power and a desire to see these subject to new entry in the interests of economic progress. It was not concerned with the pursuit of limiting cases where universal price taking (the absence of all economic power) became the aim of policy.

Policy Areas

Education

The 'big questions' of government growth and the role of the state may have formed the thematic material that permeates Peacock's work, but it was his interest in particular areas of policy for which he is best known. He had a lifelong interest in education, perhaps because, as a liberal, he saw it as a means of enabling the population to live independently, supporting human capital accumulation, and resisting the rise of dependency and the purely redistributional state. He also accepted that there were particular problems in education markets related to the public benefits associated with a well-educated population and the transactional hazards encountered in educational finance.

His paper with Jack Wiseman (1964), 'Education for democrats' advocated the development of a voucher scheme that would give parents the financial resources to purchase educational services from competing suppliers. On the demand side it was opposed by those who thought that parents could not be trusted to make such important decisions on behalf of their children, and on the supply side it was challenged by an educational establishment that, like all producer interests, preferred its customers not to have access to possible alternatives. State paternalism and state supply were to win the day for many years to come and Peacock moved away from any advisory capacity in Liberal party circles.

He did not withdraw from the debate however, and in 1978 took up a post at the University College at Buckingham, an institution that rejected direct government finance and relied upon student fees. He was Principal of the College (1980–83) and upon achieving a Royal Charter became the first Vice-Chancellor of the University. At some professional risk to himself, Peacock signalled his view that financial support for human capital accumulation in higher education should be directed through the students and not be assigned by bureaucratic and centralised formulae direct to the institutions themselves. Given the introduction of student loans in recent years he could at least reflect on the advance of this principle in higher, if not in primary or secondary, education.

Devolution

Peacock's liberalism showed itself consistently in his continuing defence of devolved decision making – if possible to the level of the individual. He was a member of the Royal Commission on the Constitution (the Kilbrandon Commission) that reported in 1973. Essentially, the majority report recommended directly elected Scottish and Welsh assemblies with specified devolved powers – a model that was eventually adopted in the Scotland Act 1998 and the Government of Wales Act 1998. With Lord Crowther-Hunt, Peacock wrote a Memorandum of Dissent to the 1973 report advocating greater decentralisation using a federal model closer to the German system. In addition to assemblies for Scotland and Wales there would be five further assemblies for regions in England – all with more substantial powers than those recommended by the majority report.

Amongst the arguments motivating the note of dissent Peacock particularly emphasised two. As he put the case 25 years later (1997a, p. 267) 'I still hold that partial devolution is an essentially unstable position'. The English regions would not long tolerate the greater influence exerted by devolved Scottish or Welsh governments. Of even greater importance, however, was his view that devolving the existing (and over-extended) functions of government would do little to empower individuals. Another political and bureaucratic layer of government controlling expenditure flows might simply empower the new representatives and officials and embolden local rent-seeking behaviour. 'The possibility was not considered that some of these functions might be carried out in different ways . . . or might be returned to the private sector' (Peacock 1976, p. 217).

Pensions

The same theme of encouraging personal responsibility and choice is revealed in Peacock's role as a member of the Inquiry into Provision for Retirement in 1984. Although he accepted that the economist as a technician should not have a privileged role in setting policy objectives, he also took the view that the clear articulation of the normative principles underlying policy was essential to any rational consideration of alternatives (1992b). When committee members were invited to list their own preferred criteria for pensions policy, Peacock (1997a, p. 320, 1997b) specified a system compatible with consumer sovereignty, based upon a minimum required standard for pensions, permitting personal saving to meet the standard, encouraging low transactions costs and, finally, conducive to the mobility of capital and labour.

Once more, Peacock's willingness to extend choice in this area by encouraging personal and portable pensions above a minimum standard met with resistance, although the committee agreed on a proposal to abolish the State Earnings Related Pension Scheme (SERPS). This was opposed by the Treasury at the time, although a highly complex system of 'contracting out' was devised. SERPS was eventually abolished in 2002 and greater flexibility and consumer choice have characterised the reform of retirement provision in the second decade of the 21st century.

It is pertinent to note that personal payments by the relatively poor into pensions or education would be assisted by the state, and Peacock was attracted by the use of finance raised at death. Here he followed J. S. Mill in favouring a tax structure that encouraged wide dissemination of estates and taxation according to the circumstances of the beneficiary rather than the donor. Peacock and Rizzo (2002) also discussed the proposal of the Italian writer Rignano to tax inherited property at progressively higher rates on successive transfers at death. Many of these ideas Peacock considered impractical, but the aim was clear – 'to produce greater equality in the distribution of wealth for the purpose of giving individuals the means to investment according to their own assessment of their welfare' (Peacock 2010, p. 137).

Broadcasting and the Arts

In 1986 Peacock was appointed Chairman of the Committee on the Financing of the BBC. It was widely assumed that, as a noted economic liberal, he had been appointed to recommend the abolition of the licence fee and its replacement by advertising revenue. In fact the report (Home Office 1986) concentrated more on the development of competition in the production of public service content and the facilitation of subscription and pay per view services that rapid technological change was making possible. Important recommendations for example were that the more popular BBC Radio 1 and Radio 2 stations should be privatised and that both ITV and the BBC should source at least 40 per cent of their output from independent producers. The licence fee was to be retained for a further period of time, but capped and indexed to inflation.

In his later writing (1997a, 2004) Peacock's suspicion of state-sponsored monopoly and his support of the normative principle of consumer sovereignty led him to oppose the channelling of licence fee revenue exclusively through the BBC. This in no way implied that he rejected out of hand the case for public service broadcasting. Various educational and other cultural objectives might lead to a case for assistance to certain types of output. He therefore recommended (2004, p. 45) that public service broadcasting should be financed through a hypothecated 'licence fee' replacement. The allocation of the available budget would then be overseen by a new council with representatives of viewers and listeners making up half of its membership. Crucially, bids for resources could be made not merely from the main terrestrial channels, but also from cable and satellite channels and for a range of purposes – including making marginal changes to the quality or accessibility of particular projects. Peacock argued that such a system would be more conducive to new entry, technical progress and 'workable competition'. The system by which the BBC received licence fee revenue and a few other broadcasters with public service requirements were regulated separately harked back to an era when spectrum scarcity severely curtailed the number of suppliers and technical non-excludability made it hard to charge for services in the market.

This liberalisation of public service broadcasting did not imply privatisation of the BBC and its transformation into a public limited company. Apart from any other consideration, property rights in past programmes financed by generations of licence fee payers would give a powerful competitive advantage to the BBC. The international reputation of the BBC in areas such as news and current affairs was also an asset that required protection. Peacock suggested that a preferable route would be to transform the BBC into a non-profit-making corporation similar to the National Trust.

To understand Peacock's approach to broadcasting it is necessary to bear in mind that he saw 'public service' elements as inextricably linked to related issues such as education, the preservation of the cultural heritage and the encouragement of the arts more generally. As noted above, he supported the idea of vouchers in education and extended this idea to consider the possibility of vouchers for museums, art galleries, concerts, libraries and a whole range of other 'cultural' activities (1993, pp. 122–8). He speculated that technical advance might even permit consumers to access a certain quantity of public service programming within a competitive environment in which pay per view television became the norm.

In general Peacock preferred subsidies to go to people rather than organisations, and this led him into continual conflict with vested interests in the world of the arts, culture and the media. In a chapter entitled 'How to lose friends and alienate people', Peacock (1993) describes his membership of the Arts Council of Great Britain and Chairmanship of the Scottish Arts Council 1986–92. He entirely endorsed John Maynard Keynes's view that the education (and hence preference formation) of the public was the main objective, and that subsidies to single companies 'were only temporary devices, rather like research and development expenditure, to give them a start in life' (1993, p. 118). Giving people the ability to access cultural events at subsidised prices would help, as experience accrued, to change preferences and lead to expenditure on the arts becoming self-sustaining.

This view of arts subsidies simply reflected his approach to the welfare state as a whole. 'The true function of the welfare state is to create the circumstances which render it unnecessary' (2010, p. 140). Hard to identify in particular policy areas and, in the eyes of many, doomed to fail in the face of the realities of democratic politics and public choice, this youthful statement of the classical liberal ideal nevertheless provides the key to Peacock's entire corpus of work in the economics of public policy.

See Also

- ▶ Art, Economics of
- ▶ Culture and Economics

- ▶ Public Choice
- ▶ Public Finance
- ▶ Welfare State

Selected Works

1949. The national insurance funds. *Economica*, New Series 16(63): 228–242.

1950. *Reform of income tax and social security payments*, Liberal Party Yellow Book.

1952. *The economics of national insurance*. William Hodge and Co.

1954a (with F. W. Paish). Economics of dependence (1952–82). *Economica*, New Series 21(84): 279–299.

1954b (with H. C. Edey). *National income and social accounting*. London: Hutchinson University Library.

1958 (with R. A. Musgrave). *Classics in the theory of public finance*. London: Macmillan.

1961 (with J. Wiseman). *The growth of public expenditure in the United Kingdom*. Princeton: Princeton University Press.

1964 (with J. Wiseman). *Education for democrats*. Hobart Paper 25. London: Institute of Economic Affairs.

1972a (with C. K. Rowley). Pareto optimality and the political economy of liberalism. *Journal of Political Economy* 80(3): 476–490.

1972b (with C. K. Rowley). Welfare economics and the public regulation of natural monopoly. *Journal of Public Economics* 1(2): 227–244.

1975 (with C. K. Rowley). *Welfare economics: A liberal restatement*. York Studies in Economics, Martin Robertson.

1976. The political economy of the dispersive revolution. *Scottish Journal of Political Economy* 23(3): 205–219.

1978a. The problem of public expenditure growth in post-industrial society. In *Post-industrial society,* ed. B. Gustafsson, 105–117. London: Croom Helm. Reprinted in Peacock (1979).

1978b. The economics of bureaucracy: An inside view. In *The economics of politics*, Readings 18. London: Institute of Economic Affairs.

1981 (with F. Forte) (eds.). *The political economy of taxation*. Oxford: Basil Blackwell.

1984. (ed.). *The regulation game: How British and west German companies bargain with government*. Oxford: Basil Blackwell.

1985 (with F. Forte) (eds.). *Public expenditure and government growth*. Oxford: Basil Blackwell.

1986 (with M. Ricketts). Bargaining and the regulatory system. *International Review of Law and Economics* 6: 3–16.

1989a (with H. Willgerodt) (eds.). *Germany's social market economy: Origins and evolution*. London: Macmillan.

1989b (with H. Willgerodt) (eds.). *German neo-liberals and the social market economy.* London: Macmillan Press.

1992a. *Public choice analysis in historical perspective*. Raffaele Mattioli Foundation, Cambridge University Press.

1992b. The credibility of economic advice to government. *Economic Journal* 102: 1213–1222. Reprinted in Peacock (1997a).

1993. *Paying the piper: Culture, music and money.* Edinburgh University Press.

1997a. *The political economy of economic freedom*. Cheltenham: Edward Elgar.

1997b. The future scope for self reliance and private insurance. In *Reforming the welfare state,* ed. H. Giersch. Berlin: Springer.

2002 (with I. Rizzo). The diffusion of economic ideas: The Rignano example. *Revista di Diritto Finanziario e Sciencza delle Finanze* 4: 547–574.

2003. *The enigmatic sailor.* Caithness: Whittles Publishing.

2004. *Public service broadcasting without the BBC?* Occasional Paper 133. London: Institute of Economic Affairs.

2008 (with I. Rizzo). *The heritage game: Economics, policy, and practice*. Oxford: Oxford University Press.

2010. *Anxious to do good: Learning to be an economist the hard way.* Exeter: Imprint Academic.

Bibliography

Beveridge, W.H. 1942. *Social insurance and allied services*. Cmd.6404. London: HMSO.

Breton, A. 1974. *The economic theory of representative government*. London: Macmillan.

Buchanan, J.M., and G. Tullock. 1965. *The calculus of consent*. Ann Arbor: University of Michigan Press.

Command Paper 5460-1. 1973. *Royal commission on the constitution 1969–73*, vol. 2. London: HMSO.

Downs, A. 1957. *An economic theory of democracy*. New York: Harper and Row.

Friedman, M., and R.D. Friedman. 1962. *Capitalism and freedom*. Chicago: University of Chicago Press.

Hicks, J.R. 1939. The foundations of welfare economics. *Economic Journal* 49(196): 696–712.

Home Office. 1986. *Report of the committee on the financing of the BBC*. Cmnd.9824. London: HMSO. (The Peacock Report).

Kaldor, N. 1939. Welfare propositions of economics and interpersonal comparisons of utility. *Economic Journal* 49(195): 549–552.

Keynes, J.N. 1891. *The scope and method of political economy*, 1st ed. London: Macmillan.

Lindahl, E. 1919. Just taxation – A positive solution. Trans. E. Henderson. Reprinted in Musgrave and Peacock (1958), 168–176.

Niskanen, W.A. 1971. *Bureaucracy and representative government*. Chicago: Aldine.

Olson, M. 1974. *The logic of collective action: Public goods and the theory of groups*. Harvard: Harvard University Press.

Puviani, A. 1903. *Teoria della Illusione Finanziaria*. Palermo: Sanaron.

Samuelson, P.A. 1954. The pure theory of public expenditure. *Review of Economics and Statistics* 36(4): 387–389.

Sen, A. 1970. The impossibility of a Paretian liberal. *Journal of Political Economy* 78(1): 152–157.

Tocqueville, Alexis de. 1965. *Democracy in America*. Oxford: Oxford University Press. Originally published 1835 and 1840.

Wagner, A. 1883. Three extracts on public finance. Translated and reprinted in Musgrave and Peacock (1958), 1–15.

Watson, G. (ed.). 1957. *The unservile state: Essays in liberty and welfare*. London: George Allen and Unwin.

Wicksell, K. 1896. A new principle of just taxation. Trans. J. M. Buchanan. In Musgrave and Peacock (1958), 72–118.

Peak-Load Pricing

Jack Wiseman

Abstract

The present-day theory of peak-load pricing is concerned with the identification of the optimal pricing structure for a particular class of products: essentially, commodities the demand for which is episodic (variable by time of day, season, or whatever) and whose technical conditions of production make storage difficult and/or are discontinuous in terms of volume.

The present-day theory of peak-load pricing is concerned with the identification of the optimal pricing structure for a particular class of products: essentially, commodities the demand for which is episodic (variable by time of day, season, or whatever) and whose technical conditions of production make storage difficult and/or are discontinuous in terms of volume.

An adequate exposition of the nature and problems of the literature concerned with this specialized problem requires that it be placed in historical perspective. There are two convergent lines of thought. The direct one originates in the discussion by Dupuit of the pricing of such services as those of bridges, Marshall's discussion of the case for the subsidization of diminishing average cost industries, and the questioning by Hotelling of the pricing of products such as railway services by the principle that 'each tub should stand on its own bottom'. For present purposes, it is enough to point out that one attempt to solve the central dilemma, identified as the need to relate price to marginal cost but yet meet all relevant opportunity costs, has been the use of multi-part pricing (a) 'standard charge' to meet overhead costs, and a 'per unit' payment (price) related to volume of consumption). Peak-load pricing can be seen as a further refinement of this line of thought, in that it seeks to relate prices to variations in the opportunity-cost situation of the producer at different points (e.g. time-periods) of consumption.

The second, distinct but clearly related historical development was the debate concerning collectivist economic planning. The question addressed in this debate is whether in a liberal collectivist economy (one in which capital is not privately owned but which is concerned to satisfy consumer choices), resources can be efficiently allocated between uses. The solution propounded is that the absence of a market in capital can be overcome by instructing all producers to equate marginal costs and prices, so replicating the outcome of perfectly competitive markets in a private capitalist economy. The theory of multipart and peak-load pricing can be seen as an extension of this specification of the optimum allocative conditions to the special circumstances of the specified industries, in which the 'competitive' solution is argued to be excluded by the technical conditions of production. It follows from this specification that the problem (and theory) has relevance irrespective of the means being adopted to allocate resources elsewhere in the economy. If the relevant technical conditions pertain, a separate 'solution' is needed whether or not competitive markets or other means are utilized to seek optimum allocative efficiency elsewhere.

The theory of peak-load pricing can be discussed in two related contexts: a narrow one concerned with the technical characteristics of the problem within the defining constraints explained above, and a much broader one concerned with the relation between this narrow specification and the wider questions raised by the theory of public utility pricing and collectivist economic planning. This in turn will provide the material for a concluding evaluation of the current state, policy-relevance, and political development of the theory of peak-load pricing.

It is of interest that the specialist literature concerned with peak-load pricing has come to be preoccupied with what are seen as essentially practical problems. Writers are concerned to develop the theory of marginal cost pricing in ways that will contribute to the specification of optimal tariff structures for the supply of actual products which are perceived to have) 'peakload' characteristics: there is a whole sub-literature, for example, concerned with electricity prices. A useful introduction to this 'narrow perspective' literature, providing also ample further references for the interested reader, is provided in a Symposium in *The Bell Journal of Economics* (1976). The introductory article by Joskow (1976) in this Symposium provides a useful taxonomic survey. He distinguishes three broad approaches, labelled 'somewhat artificially' the American, British and French. All three approaches find their intellectual heritage in the theory of public utility pricing, which is to say that the writers concerned would accept, as the hallmark of an 'efficient' pricing and investment structure, conformity with the 'optimum conditions of choice' of mainstream welfare economics (Joskow 1976, describes an) 'implicit social welfare function' in a fashion that exemplifies this). They are distinguishable rather by their emphasis and coverage, which in turn it is perhaps not too fanciful to relate to the typical form of national industrial organization: regulated private monopoly in the USA, nationalized industry in the UK, state enterprise within an indicative plan in France.

The American approach specifies an enterprise with homogeneous productive capacity. If the location and magnitude of the peak is not affected by the relative prices charged in peak and off-peak periods, then the efficient solution is a two- (or multi-) part tariff which separates capacity and operating costs and allocates the marginal opportunity-cost of capacity to peak users. That is, off-peak users pay marginal operating costs, but peak users also pay marginal capacity costs. This solution requires that the relevant physical capacity be not indivisible in the relevant respects.

If these conditions are not met (if demand at peak or off-peak is sensitive to relative prices: if the off-peak load utilizes plant to capacity when charged at short-run marginal cost), then a problem of 'shifting peaks' emerges, and the identification of the efficient tariff becomes much more complex. Detailed information is needed about both production costs and about the character and inter-relationships of period demand-functions. This information is the input requisite for a formal ('welfare-optimizing') solution which maximizes net benefits (total revenue plus consumers' surplus less opportunity costs) and needs further to be supplemented by the assumption (value judgement) that individual gains and losses can be weighted equally.

The British literature is concerned with a more sophisticated technology: it postulates a mix of units of physical capital with different investment and operating costs and maintenance requirements. The purpose is to determine the optimum utilization of the capital stock by identifying the long-run marginal costs relevant to the fixing of prices to meet a fixed set of consumer demands.

Peak and off-peak periods are determined by the variations in both demand and in the type of capacity available. The analysis is more sophisticated (and realistic) than the American in its handling of the technology of production, but less satisfactory in dealing with (or, rather, failing to incorporate) the problems of shifting peaks.

The French contribution comes largely from economists with industrial rather than university academic affiliations. This is very much in the Dupuit tradition, which in the present instance has given technically sophisticated economists considerable practical influence over electricity pricing and investment policy. This is reflected in a concern for operational relevance. The defining characteristic of their contribution is that it attempts to develop the American and British approaches in ways that facilitate the practical application of the constructs to the actual determination of electricity prices. Joskow (1976) identifies three important contributions.

First, they are concerned that demand may be uncertain as well as periodic, with the implication that (unforeseen) demand may exceed current capacity and/or standby capacity must be kept available. Second, any curtailment of supply because of uncertainty may itself impose costs on both the demand and the supply side, and these 'marginal curtailment costs' have to be seen as a further element in the determination of the efficient pricing-and-investment plan. Third, an efficient tariff structure must embrace all opportunity-costs incurred up to the point of consumption. This is no trivial matter: the question has traditionally been seen as concerned with the relation between a production supply technology and a variable consumption demand. But in the case of electricity, for example, the distribution system may account for around half of total costs. The pricing structure thus needs to integrate a centralized generation system with its related technological problems and diversities, and a dispersed distribution network directed to the satisfaction of the disaggregated consumption demands of individuals and small groups.

The literature under review is most satisfactory when it is concerned with what is essentially an 'engineering' problem, and becomes progressively less persuasive as we move from this restrictive context towards the requirements of a 'political economy' model bearing on economic (pricing) policy in the real world. Even at the restricted level, our survey suggests that there are significant outstanding problems to be solved: the complementary insights of the British and American approaches, for example, await integration in a common construct that can incorporate both technological diversity and shifting peaks. More important, the writers themselves recognize that the 'engineering' approach can be translated into pricing policy recommendations only if a solution can be found to related valuation problems which tend to be dealt with by essentially arbitrary assumptions. There is recourse to such concepts as consumer surplus and opportunity-costs, value judgements about the weighting of gains and losses, and so on, with little recognition of the essentially subjective nature of these concepts, and the difficulties that this creates for the translation of an objectively specified (engineering) formulation into a system of pricing rules/procedures. The French approach, with its emphasis on uncertainty, underlines these difficulties. How to write into a peak-load pricing model, for example, individual judgements as to the possibility that within the next decade someone will invent an efficient generator that can be housed in the family basement?

The contribution and limitations of the peak-load pricing literature can be most easily summarized by relating the description so far to the broader streams of thought referred to at the outset: those of public utility pricing and of collectivist economic planning. In respect of the former, the peak-load literature does nothing to resolve the fundamental logical problem that arises even within the welfare economics model of which that literature is a special part. That is that *all* scarce resources have opportunity-costs over an appropriate period of time; that is the implication of scarcity. If prices reflect all these opportunity costs, then the problem of conflict between average and marginal costs disappears: there will be as many (opportunity-cost) prices as there are relevant time-periods. If the price structure differs from this (which is the conclusion of the ‘orthodox’ analysis), then that is because some relevant opportunity-costs are being ignored. The) ‘deficit’ that this is argued to imply must be borne by someone, and the decision about how it ‘should’ be borne is effectively a policy (distributional) value judgement about which the economist *qua* economist can have nothing to say (Wiseman 1957). The peak-load pricing literature contributes nothing to the resolution of this prior problem.

An even more fundamental difficulty surfaces when the literature is placed in the broader, economy-wide context of the collectivist economic planning debate. Essentially, the liberal collectivist solution to the resource-allocation ‘problem’ consists in the replacement of the competitive search for profit in the capitalist economy by obedience to a ‘cost rule’ (‘make marginal cost equal to price’) in the collectivist one. What is identified as a characteristic of the *outcome* of the capitalist process becomes a rule of cost-behaviour in the collectivist one. For the cost-rule to be operationally plausible, it is necessary that the costs concerned should be objective, since otherwise the instruction to make marginal cost equal to price amounts simply to an exhortation to the decision-taker to choose the plan he believes ‘best’. Effectively, *any* pricing and investment decision could be made to appear consonant with such an instruction: and if the *content* of the marginal cost) ‘plan’ is specified for the manager, then the effective allocation decision is simply shifted to those who give the instructions, and their subjective evaluations then determine the content of) ‘marginal cost’.

In briefest summary of a complex literature, once it is recognized that the opportunity-costs that determine resource-allocations through time are the subjective evaluations of future possibilities of those who make the relevant decisions, there can be no ‘objective’ way to determine prices by reference to costs. The money outlays into which opportunity-costs are commonly translated must in a policy context be predicted future money outlays, and so must be someone’s opinion rather than have ‘objective’ or ‘scientific’ status (Wiseman 1983). If this is true of costs (and cost-rules) generally, it must be true equally of those particular costs that are the concern of the multi-part pricing literature.

The literature is not without interest or potential practical utility, provided that it is treated as no more than a particular input to the subjective opportunity-cost plans bearing on a special set of products. There are incidental recognitions of this, and it is clearly a plausible inference from the French development of the problem. But until there is an explicit recognition of the need to integrate the technological (‘engineering’) characteristics into a subjective opportunity cost framework, any claim that the literature enables the identification of economically or socially-efficient pricing structures must be treated with caution.

See Also

- ▶ Marginal and Average Cost Pricing
- ▶ Public Utility Pricing

Bibliography

Joskow, P.L. 1976. Contributions of the theory of marginal cost pricing. *Bell Journal of Economics* 7(1): 197–206.

Symposium on Peak Load Pricing. 1976. *Bell Journal of Economics* 7(1): 195–248.

Wiseman, J. 1957. The theory of public utility price – An empty box. *Oxford Economic Papers* 9: 56–74.

Wiseman, J., eds. 1983. *Beyond positive economics?* London: Macmillan.

Peasant Economy

N. H. Stern

Keywords

Credit; Green Revolution; Land reform; Peasant economy; Rational behaviour; Rural–urban migration; Sharecropping; Urban unemployment

JEL Classifications

O1

A peasant is someone who lives in the country and works on the land (the word derives from the French *paysan*). Taking this definition, the topic 'peasant economy' concerns the analysis of the economic decisions and interactions of peasants, their relations with other agents and the rest of the economy, the determinants of the general level and distribution of their economic welfare, and how their position might move over time or be affected by policy. As such it is very broad in scope, involving the study of the economic life of around half the world's population. The term 'peasant' is sometimes used in a somewhat narrower sense in economics to mean the small farmer (tenant or smallholder) as opposed to the agricultural labourer or very large landowner. The peasant economy would then be one where farming was conducted mainly by tenants and smallholders. Even under this narrower definition it is clear that vast numbers of individuals are included.

There are fundamental differences amongst economists in their views of the way in which the peasant economy functions and these underlie many of the strong disagreements over policy. The main sources of the differences concern views on the 'rationality' of economic behaviour by individuals, the competitiveness and efficiency of markets, the importance and implications of the distribution of power and wealth, and the role of institutions, cultures and beliefs. Whilst we cannot provide a detailed description of these general views we shall try to give a flavour of their diversity and focus on the basis of their differences. We shall then examine some specific issues and problems including the objectives of peasants and others in terms of profit, utility and attitudes to risk; labour markets; credit; sharecropping; and relationships between size of holding and productivity. Concentration will be on the literature since the Second World War, although many of the issues concerned and divided some of the outstanding economists of the 19th and early 20th century.

One of the most clearly stated and definite views places the peasant economy firmly within the standard competitive analysis; see for example Schultz (1964). Within the constraints of their knowledge, it is argued, participants in the peasant economy make the best use of the assets available to them. Each agent makes production, working and spending decisions to maximize utility or profit. This is essentially the notion of rationality in this context: individuals have preferences and act according to them. Markets for labour, land, credit, inputs and outputs, consumer purchases and so on function competitively and efficiently. The outcome for the usual reasons is therefore a Pareto efficient allocation. The role of policy is then to improve knowledge, increase assets and, if desired, to improve the distribution of income.

At another extreme we find the views of those such as Myrdal (1968), who believes that markets and prices play a minimal role. He argues that few people calculate in terms of costs and returns and that, even if they do, such calculations are not the primary determinants of their behaviour. Further, he argues that many transactions are not of the market type at all, and where markets do exist they are very far from perfect. He pleads for an institutional analysis of behaviour and the workings of the economy. Further, he suggests direct controls to implement policy; he calls these non-discretionary controls as opposed to the manipulation of prices, where individuals are left to take their own decisions.

Away from these extremes we have varying emphases on the role of rational behaviour, incentives and market structure. For example, Lewis (1955) regards institutions, legal structures and political and religious attitudes and practices as major determinants of the form of incentives. Thus he suggests that land reform may be a prerequisite to successful agricultural extension if, without it, farmers believe that others will reap the fruits of their improvements. Since the 1970s there has been substantial concentration on the forms of peasant arrangements for cultivation, the incentives which they give and the reasons for their selection. A central example (discussed briefly below) has been the study of sharecropping following the questions and analysis of Marshall in Chapter X, Book VI, of his *Principles of Economics*. In this context individuals are seen as rational but face problems of information and supervision in designing and implementing agreements for the use of land, labour and other inputs.

Marxist writers have emphasized property and power. For example, Bhaduri (1973) suggests that landlords manipulate indebtedness over their labourers and tenants to maintain a very tight hold over their freedom. He argues from his model that landlords have an incentive to block technical change and that progress requires expropriation.

These views are generalizations about the world and no single study could provide a conclusive test between them. An empirical judgement should be based on the accumulated experience of detailed studies. Here economists have not been as active as perhaps they should in conducting economic studies of peasant societies to examine how the theories they are discussing fare in the field (compare the many studies by anthropologists; see, for example, Srinivas 1960 and 1976, and Wiser and Wiser 1971). Nevertheless, many studies are available (see for example, Bailey 1957; Epstein 1962; Haswell 1975; Bell 1977; Bliss and Stern 1982, and for further references Binswanger and Rosenzweig 1984; see also the bibliographies of village studies prepared at the Institute of Development Studies, Sussex – Lambert 1976 and 1978). One should not perhaps expect a clear, single picture to emerge; people and societies vary considerably. However, it seems that neither of the simple descriptions of Myrdal and Schultz are remotely adequate as generalizations. The institutional structure and conventions concerning the disposition of land and labour (for example, the form of ownership and duties of owners, structure of tenancy agreements, restrictions on the obligations of labourers and so on) will be of considerable importance in determining cultivation decisions. Individuals vary greatly in their ability to make the most of their circumstances. Nevertheless, most of the studies point to strong economic responses and these are often rapid and subtle: the Myrdal picture is clearly unacceptable.

We comment briefly on some of the particular positive issues that have been prominent in theory and applied work. The objectives of peasants have been modelled in terms of profit and utility and in varying ways relative to uncertainty (for an early discussion see Chayanov 1925). Thus, for example, Hopper (1965) suggests that simple maximization of expected profit provided a good description of farming decisions in the village he studied in North India. This seems implausible in a poor society and for a risky activity, and a number of models of behaviour under uncertainty have been considered. These include the standard model of expected utility maximization and 'survival algorithms' where individuals attempt to minimize the probability of falling below 'disaster level'. The implications can be very different from simple profit maximization. Under expected utility maximization with risk aversion the expected value of the marginal product of an input would, in equilibrium, be above the price of the input (possibly well above) whereas with profit maximization we must have equality (see, for example, Bliss and Stern 1982).

Two central issues in discussion of the labour market have been, first, the relationship between wages and the marginal product of labour and, second, migration. On the former some appear to have argued that the marginal product is zero. This receives little empirical or theoretical support in that an extra hour of work in agriculture usually has some contribution to production. The question of whether the withdrawal of an extra person from agriculture reduces output and by how much depends on the response to the departure by others. Whilst the marginal product of an hour or day is unlikely to be zero, it is quite possible that it may be less than the wage in the case of family labour where there are perceived costs in working for others or of hiring labour (see for example Sen 1975).

Migration decisions have been examined extensively, both in theory and practice, in terms of expected differences in net incomes or utility from making a move. Of particular influence was the paper by Todaro (1969) in which he proposed a model where the probability of employment in the town was equal to the number of jobs divided by the number of seekers. If rural and urban wages and urban employment are fixed, the number of seekers adjusts to make, in equilibrium, the expected urban wage equal to the rural wage. If we associate the job seekers with the employed plus the unemployed then this is a theory of urban unemployment with the striking implication that an increase in the number of urban jobs increases unemployment. The model has been extended, elaborated and tested by many authors (see, in particular, Fields 1975; Sabot 1982; Todaro 1976).

The role of credit, for example, the much easier access and cheaper rates available to the richer farmers (Griffin 1974) and its use in manipulation and control (Bhaduri 1973) have been major issues. It is an area where data are particularly difficult to collect and good empirical studies are rare (a notable exception in the context of fishing is Platteau et al. 1985).

Share-cropping was discussed carefully by Marshall in his *Principles*. Following the book by Cheung (1969), it has become a popular issue in recent research. Cheung contrasted his view of sharecropping as an efficient arrangement (with the tenancy contract clearly defined to stipulate inputs) with that of Marshall, who had pointed to the possibility that the tenant who receives half the output may not push the level of an input as far as someone who receives the full amount of the marginal product. Many of Cheung's arguments were, however, anticipated by Marshall in his account which contains a description of how the landlord might try to enforce higher input levels. More recently attention has been focused on sharecropping as a means of sharing risk between landlord and tenant and as providing incentives for the tenant which would not be present under simple wage labour (see Binswanger and Rosenzweig 1984, for references).

The proposition that larger holdings may have lower output per acre has been the subject of much theoretical and empirical discussion. In Indian studies it receives more support for comparisons across districts than within villages. Possible reasons for the phenomenon, where it occurs, include more labour input per acre on smaller family plots (where labour may be applied beyond the point where the marginal product is equal to the wage) and faster population growth (and thus greater subdivision of holdings) on fertile land. For further discussion, see Sen (1975).

On the policy side some of the major issues have been land reform, the dissemination of technical change, the pricing of output and the supply and pricing of crucial inputs such as water, fertilizer and draught power. We shall be very brief since our main emphasis has been on the functioning of the peasant economy. Land reform in the sense of redistribution has been very difficult to achieve, in part because many of those who have it will make great efforts to resist losing it. It has sometimes been argued that the (supposed) inverse relationship between size of holding and land productivity will imply that a more egalitarian distribution of land will yield higher total output. Agricultural extension has long been seen as part of government policy, but it has become particularly prominent with the arrival of the newer varieties of seeds (the so-called 'Green Revolution') which are particularly responsive to water and fertilizers. Of special concern has been the differential impact of the advances on different groups in the population and how the changes might be influenced to provide greater benefits to the poor.

The relative price of food and the implicit or explicit taxation of peasants have been seen as critical aspects of the availability of food (and its price) to the rest of the economy as well as influencing growth within and outside peasant agriculture. Much turns on the assumed elasticity of response. A further important feature of government policy concerns the pricing and supply of inputs. The effects on agricultural production and on the welfare of peasants and labourers can be substantial, the most obvious example being irrigation.

The study of the peasant economy is a subject for which careful economic theorizing is critical since transactions can have special structures, uncertainty will be central, and economic relations will be strongly influenced by institutional arrangements. And those theories should be tested against, and arise from, detailed empirical observation since the successful application of the theories turns on which of the structures are relevant for the particular peasant economy under examination.

See Also

- ▶ Agriculture and Economic Development
- ▶ Peasants
- ▶ Sharecropping

Bibliography

Bailey, F.G. 1957. *Caste and the economic frontier*. Manchester: Manchester University Press.

Bell, C.L.G. 1977. Alternative theories of share-cropping: Some tests using evidence from North-East India. *Journal of Development Studies* 13: 317–346.

Bhaduri, A. 1973. Agricultural backwardness under semi-feudalism. *Economic Journal* 83: 120–137.

Binswanger, H.P., and M.R. Rosenzweig. 1984. *Contractual arrangements, employment, and wages in rural labour markets in Asia*. New Haven: Yale University Press.

Bliss, C.J., and N.H. Stern. 1982. *Palanpur: The economy of an Indian village*. Oxford: Oxford University Press.

Chayanov, A.V. 1925. *Organizatsiya krest'yanskogo khozyaistva* [Peasant farm organization]. Moscow. Trans. as *The Theory of Peasant Economy*, ed. D. Thorner, B. Kerblay and R.E.F. Smith. Homewood: Irwin (AEA Translation Series), 1966.

Cheung, S.N.S. 1969. *The theory of share tenancy*. Chicago: University of Chicago Press.

Epstein, T.S. 1962. *Economic development and social change in South India*. Manchester: Manchester University Press.

Fields, G.S. 1975. Rural urban migration, urban unemployment and underemployment, and job search activities in LDCs. *Journal of Development Economics* 2: 165–187.

Griffin, K.B. 1974. *The political economy of agrarian change*. London: Macmillan.

Haswell, M. 1975. *The nature of poverty*. London: Macmillan.

Hopper, W.D. 1965. Allocation efficiency in 'traditional Indian agriculture'. *Journal of Farm Economics* 47: 611–624.

Lambert, C.M., ed. 1976. *Village studies I*. Institute of Development Studies, University of Sussex.

Lambert, C.M., ed. 1978. *Village studies II*. Institute of Development Studies, University of Sussex.

Lewis, W.A. 1955. *The theory of economic growth*. London: George Allen & Unwin.

Marshall, A. 1920. *Principles of economics*, 8th edn. London: Macmillan, 1959.

Myrdal, G. 1968. *Asian drama: An enquiry into the poverty of nations*. Harmondsworth: Allen Lane.

Platteau, J.-P., J. Murickan, and E. Delbar. 1985. *Technology, credit and indebtedness in marine fishing*. Delhi: Hindustan Publishing Corporation.

Sabot, R.H., ed. 1982. *Migration and the labour market in LDCs*. Boulder: Westview.

Schultz, T.W. 1964. *Transforming traditional agriculture*. New Haven: Yale University Press.

Sen, A.K. 1975. *Employment, technology and development*. Oxford: Oxford University Press.

Srinivas, M.N., ed. 1960. *India's villages*. Bombay: Asia Publishing House.

Srinivas, M.N. 1976. *The remembered village*. Oxford: Oxford University Press.

Todaro, M.P. 1969. A model of labour migration and urban unemployment in less developed countries. *American Economic Review* 59: 138–148.

Todaro, M.P. 1976. *Internal migration in developing countries: A review of theory, evidence, methodology and research priorities*. Geneva: International Labour Office.

Wiser, W.H., and C.V. Wiser. 1971. *Behind mud walls, 1930–1960*. Berkeley: University of California Press.

Peasants

Keijiro Otsuka

Abstract

While traditionally peasants are regarded as subsistence-oriented, full-time, and small-scale farmers, many small-farmers are part-time farmers engaged in both cash-and food-crop farming and non-farm jobs. Therefore, peasants may be defined as small-scale, family based farmers, including both owner cultivators and tenants. A major question is whether the peasant mode of production is socially efficient. Because of the absence of scale economies, the advantage of risk sharing under share tenancy contracts, and the inefficiency of agricultural labour contracts due to the difficulty of supervision, small-scale family based farming system, including share tenancy, is a socially efficient system in low-wage economies.

Keywords

Access to land; Cash crops; Contract choice; Contract enforcement; Credit markets; Efficient allocation; Family labour; Fixed-rent tenancy; Food security; Human capital; Insurance markets; Land use rights; Market failure; Marshallian inefficiency of share tenancy;

Monitoring costs; Peasants; Reputation; Risk sharing; Scale economies; Sharecropping; Tenancy markets

JEL Classifications

O1

Defining who 'peasants' are is not an easy task for social scientists interested in rural economies and their transformation over time. The traditional image of peasants may be small-scale, full-time farmers, who have some access to land, depend largely on family labour, and produce food primarily for home consumption. The main characteristics of peasants, however, have changed over time in the process of economic development that has accompanied the penetration of markets into rural areas. Many of them are part-time farmers engaged in both farming and non-farm jobs, and produce cash crops in addition to food crops. Yet family farms continue to dominate throughout the world, contrary to the traditional view that they are remnants of feudal society and are bound to disappear as modernization proceeds (Hayami 1996). Therefore, in order for the concept of peasants to be relevant to the present world, it seems sensible to define peasants simply as small-scale, family based farmers.

Thus, we exclude agricultural labourers dependent on wage employment, and 'capitalist' farmers, large landlords, and plantation owners who operate large farms using hired labour. The major categories of peasants are owner-cultivators and tenants, and tenants can be further classified into leaseholders (or fixed-rent tenants) and share tenants (or sharecroppers). Commonly these peasants are managers of farms, engaged in multiple farm tasks such as land preparation, fertilizer application, and the supervision of wage workers hired for simple tasks such as weeding and harvesting. Tenants are subject to terms of contracts, which are often unwritten and implicit such as the careful maintenance of irrigation facilities and diligent work on assigned tasks. Being small-scale, efficient farm production for food security is a major concern in a peasant society.

We also regard small cultivators in customary land tenure areas as peasants, even though they are neither owner-cultivators nor tenants. These cultivators have the use right on land as long as they continue to cultivate it, but typically they do not possess ownership rights. Thus, for example, once land is put into fallow, the cultivator tends to lose the use right. The future use of land, as well as the inheritance of land use rights, is determined by the leader of the extended family or the village chief. Such insecurity of tenure arising from the uncertain access to land in future may reduce incentives to invest in the long-term improvement of land, because those who invest may not be able to reap the benefits in the future. Although it may appear that the same argument applies to tenancy contracts, so long as the landowner has the right to terminate the contract, it is landowners, but not tenants, who make long-term investment decisions in the case of tenancy. Thus, whether the tenure insecurity results in underinvestment in land improvement is a major empirical question particularly relevant to customary land tenure areas (Besley 1995). A critical question in the study of customary tenure institutions is whether efforts to invest in land – for example, tree planting and terracing – confer strong individualized land rights *ex post,* so as to provide proper incentives to invest *ex ante* (Otsuka and Place 2001).

Peasant farms are small because scale economies are absent under the prevailing labour-intensive farming systems, which are characterized neither by indivisibility caused by large-scale mechanization nor by the specialization and division of labour among farm workers. Thus, the optimum farm size is likely to be small. In Asia, the average size of rice-growing farm households seldom exceeds two hectares, and it can be as low as 0.5 hectares in Java, Bangladesh, and China (David and Otsuka 1994). Extremely large farms, including haciendas, plantations, and estates, were created by force by colonial governments, not by market forces. Once they are created, however, they tend to persist, even though their sizes exceed the optimum, primarily because the land sales market does not function due to imperfect credit markets (Binswanger and Rosenzweig 1986). The inverse correlation between farm size and productivity, often measured by yield per hectare, is widely observed in South Asia, which indicates the existence of scale diseconomies (Otsuka 2007). Such scale diseconomies are likely to arise from the difficulty faced by large farmers in supervising hired workers in spatially wide and ecologically diverse farm production environments.

It is widely believed that peasants are poor but efficient in resource allocation – the ‘efficient but poor’ hypothesis of Schultz (1964), which argues that investments in human capital and the dissemination of new technologies are the keys to improving their livelihood. A major challenge to the Schultz thesis is the so-called Marshallian inefficiency of share tenancy. According to this theory, a share tenant does not work as hard as an owner-cultivator or a leasehold tenant, because he receives only a fraction of the value of the marginal product of labour. Inexplicably, output sharing rate under share tenancy is fifty–fifty not only historically in France, Italy and the antebellum South in 19th century United States, but also in many contemporary developing countries (Hayami and Otsuka 1993). If sharing rate is not fifty–fifty it is two-thirds for the tenant and one-third for the landlord almost without exception. It is argued by the advocates of the Marshallian thesis that output-sharing is like the imposition of proportional income tax on a tenant, which discourages him from working hard. For this reason, share tenancy is prohibited by land reform laws in a number of countries in Asia.

It must be pointed out that Marshall himself (1890) did not necessarily support the Marshallian thesis: he pointed out the major shortcomings of this argument in a footnote, which was later elaborated upon by Johnson (1951) and Cheung (1969). The main point is that both the landlord and the tenant can be made better off by adopting a fixed-rent contract, which does not distort work incentives as the tenant receives the entire marginal product of labour, and then by sharing the larger ‘pie’ between the two parties. Marshall argued that, if the work effort of the share tenant can be monitored costlessly by the landlord, the share tenant will be forced to work as hard as a fixed-rent tenant. The implication is that share tenants tend to shirk unless they are effectively monitored or provided extra incentives to work harder. It is also generally agreed that, despite such problems, share tenancy is prevalent because of the risk sharing advantage; the production risk is shared between share tenants and landlords, unlike with fixed-rent contracts in which all the risk is shouldered by the tenants. This argument is plausible considering the absence of insurance markets and the existence of substantial production risk in poor agrarian communities.

Because of the existence of monitoring costs, share tenancy is inefficient in the literal sense of the word. One may argue, however, that since we cannot avoid monitoring costs in the real world, it is misleading to argue that share tenancy is inefficient; it is ‘second-best’ efficient, even if a tenant shirks. Unlike other areas of contract studies, there have been a huge number of empirical studies comparing yields per hectare between share tenancy and owner-cultivation or fixed-rent tenancy. According to a summary of earlier empirical studies by Hayami and Otsuka (1993) and a number of subsequent empirical studies, the difference in yield is found to be generally insignificant, suggesting that resource allocation under share tenancy is not significantly different from the ‘first-best’ efficiency. It is true that the differences in land and labour qualities are not properly controlled for in some studies, so that their yield comparisons are not as rigorous as they ought to be. However, since there is no reason to believe that share tenants are endowed with greater human capital and cultivate higher-quality land, the empirical evidence can be taken to imply that share tenancy is efficient, or at least not as inefficient as the Marshallian thesis assumes.

Hayami and Otsuka (1993) argue that significant shirking by share tenants is prevented by multifaceted, enduring personal relationships between tenants and landlords and the community mechanism of contract enforcement. More often than not, the landlord selects the share tenant, who is deeply related by kinship or community ties. Therefore, if the dishonest behaviours of a tenant are detected, he will be penalized not only by the termination of the share contract but also by the discontinuation of multifaceted personal relationships. Furthermore, he will not be able to find other landlords in the same community who are willing to offer new share contracts because of the loss of reputation as an honest and hard-working tenant. In this way shirking is prevented, which supports the Schultz thesis.

The above argument implies that if the share contract is deemed to be short term – for example, one season or one year – the share tenant is likely to shirk because, regardless of whether he shirks or not, the contract will be terminated at the end of the cropping season. Indeed, according to Hayami and Otsuka (1993), significantly lower crop yields under share tenancy are typically found in India, where large landlords rotate share tenants season after season in order to avoid the implementation of the 'land-to-the-tiller program', which attempts to transfer land to the tenant. Since the presumption of the land-to-the-tiller program is that there is a single tenant on each piece of land, its implementation becomes difficult if there are many tenants.

If share tenancy is not significantly inefficient, we should not observe the inverse correlation between farm size and productivity, because larger and less productive farmers can gain by renting out a part of their lands to smaller and more productive share tenants. Indeed, in general, the inverse correlation is seldom found in South-east Asia, where tenancy markets are generally active, whereas it is often observed in South Asia where tenancy markets tend to be suppressed or discouraged by land reform laws (Otsuka 2007).

In theory, it is considered that a fixed-rent contract will be chosen only if the tenant is risk neutral, because it provides proper work incentives to tenants who are willing to assume production risks. It is, however, highly unlikely that tenants, who are often landless and poor, do not care about the production and income risks. Although rigorous analysis is required, casual observation as well as a brief literature survey suggest that fixed-rent tenancy is more common than share tenancy in sub-Saharan Africa, unlike Asia where share tenancy is dominant. Since African farmers are poorer, it seems unreasonable to assume the risk neutrality of fixed-rent tenant farmers in sub-Saharan Africa. Indeed, they grow multiple crops presumably to diversify the production risks. One missing factor that possibly affects the contract choice is the cost of metering output under share tenancy. Since a share tenant has an incentive to under-report the amount of output to increase his share of income, the landlord must be able to meter the output effectively in order to prevent the tenant's cheating. In the case of rice farming in Asia, either the landlord himself watches the harvesting or he sends some dependable person, like his son, to the field on the designated harvesting days. The cost of watching the harvest will be high for absentee landlords and widows who have no farming experience. This is one of the reasons why such landowners usually offer fixed-rent contracts.

The importance of the cost of metering output in contract choice is illustrated by the case of the share contract of cocoa farming in Ghana, whose harvesting season lasts for more than a few months; instead of sharing output, the tenant and the landlord share the ownership of the land after the tenant finishes planting the cocoa trees (Otsuka and Place 2001). One plausible hypothesis is that a precondition for share tenancy to be adopted is a short harvesting season, so that the cost of measuring output for the landlord is reasonably low. This hypothesis may explain why share tenancy is common in Asia, where rice and wheat are the major crops, whereas fixed-rent contracts are common in sub-Saharan Africa where maize, cassava, and other food crops which are harvested for prolonged periods are the major crops. Interestingly enough, in Ethiopia, where wheat, barley, and teff (a uniquely grain grown in this country alone) are the major crops, share tenancy is common (Benin et al. 2005). It is, however, fair to say that how crop choice and contract choice are related is an important empirical question to be investigated further.

As the population pressure on limited land resources increases in developing countries, land becomes scarce and tenancy becomes important in achieving the efficient allocation of land among farm households by transferring cultivation rights from land-rich to land-poor households. Also becoming increasingly important are non-farm jobs, as small-scale farming subject to seasonality cannot ensure a decent living standard. Thus, the development of the rural non-farm economy is the norm rather than the exception throughout developing countries (Haggblade et al. 2006), in which members of small-scale family farms are engaged not only in food production but also the production of cash crops and, more importantly, in non-farm jobs (Bliss and Stern 1982; David and Otsuka 1994; Hayami and Kikuchi 2000; Quisumbing et al. 2004). Thus, the conventional characterization of peasants as self-sufficient food production units as envisaged by Chayanov (1966) is no longer valid.

This does not imply, however, that markets work competitively in rural economies, so that rural households allocate labour time among food production, cash crop production, and non-farm activities, purchase all factors of production freely so as to maximize profits, and purchase goods and services so as to maximize utility. This separability of production and consumption decisions does not hold if markets are imperfect (Singh et al. 1986). Therefore, in a Chayanovian world, production and consumption decisions must be made simultaneously.

Although many commodities and factors of production can be purchased and sold at competitive markets, there are also serious market failures. First of all, insurance markets fail to develop, as bad harvests negatively affect the income of all farmers in the locality (Binswanger and Rosenzweig 1986). Second, credit markets tend to be imperfect primarily because of the lack of collateral, except for owner-cultivators who can use land as collateral. Third, labour markets, in general, do not function efficiently because of the difficulty in labour supervision. Thus, the labour market is typically thin or hired labour is employed only for such simple tasks as weeding and harvesting, activities which can be monitored easily (Hayami and Otsuka 1993). If hired labour is employed for tasks which require care and judgment, such as water management, land preparation, and fertilizer application, the farm operation becomes inefficient. This is likely to be the main reason for the inverse correlation between farm size and productivity, in view of the fact that the suppression of land tenancy transactions forces large farmers to employ seasonal labour, often called ‘permanent’ labour, for tasks requiring care and judgement in South Asia.

If the labour market fails, the response of peasants to new marketing opportunities and new technologies can become sluggish or even perverse (de Janvry et al. 1991). For example, when the price of cash crops increases, their supply may not increase much, because farmers must depend solely on family labour without employing additional hired labour, which ought to be available at constant wage rates in the presence of a competitive labour market. Similarly, technological change in the food sector may not lead to large increases in the market supply of food if labour markets fail and food markets do not function effectively.

In view of the increasing involvement of peasants in market transactions, it is critically important to strengthen the efficiency of marketing sectors through investing in roads, communication facilities, and marketing information, such as the establishment of quality standards for farm products, in order to improve their well-being. According to Hayami (1996), peasant entrepreneurs significantly contributed to the development of rural commerce and industries in the process of economic development in East Asia.

It must be emphasized that, given the difficulty in labour supervision, we can hardly expect the farm labour markets to function efficiently. In all likelihood, it is more realistic to promote efficient tenancy transactions, be it share or fixed-rent tenancy, if we hope to develop peasant sectors in the rapidly globalizing world where markets penetrate increasingly into rural areas. Efficient tenancy markets will increase the responsiveness of peasant sectors to new market and technological opportunities by facilitating the reallocation of land from households endowed with meagre family labour relative to land to those with an abundant supply of family labour.

The importance of tenancy transactions will continue to increase as an economy develops further. An efficient farm size expands with an increase in the wage rate, which makes it profitable to introduce large-scale mechanization to save labour. The traditional peasant mode of labour-intensive production on small farms, therefore, will no longer be sustainable in high-wage economies. Because of the scale economies associated with large-scale mechanization, viable farmers accumulate large cultivation areas through land tenancy. The practical question is at what farm size we can legitimately claim that farmers are no longer peasants. Although a clear and unanimously acceptable answer can hardly be given, I would like to propose that the issue of peasants ceases to be relevant when the issue of food insecurity associated with small farm size is resolved through farm size expansion, as well as the development of efficient marketing systems and technological changes in food production.

See Also

- ▶ Access to Land and Development
- ▶ Land Markets
- ▶ Peasant Economy
- ▶ Sharecropping

Bibliography

Benin, S., M. Ahmed, J. Pender, and S. Ehui. 2005. Development of land rental markets and agricultural productivity growth. *Journal of African Economies* 14: 21–54.

Besley, T. 1995. Property rights and investment incentives. *Journal of Political Economy* 103: 913–937.

Binswanger, H.P., and M.R. Rosenzweig. 1986. Behavioral and material determinants of production relations in agriculture. *Journal of Development Studies* 22: 503–539.

Bliss, C.J., and N.H. Stern. 1982. *Palanpur: The economy of an Indian village*. Oxford: Oxford University Press.

Chayanov, A.V. 1966. *The theory of peasant economy.* Homewood: Irwin.

Cheung, S.N.S. 1969. *The theory of share tenancy.* Chicago: University of Chicago Press.

David, C.C., and K. Otsuka. 1994. *Modern rice technology and income distribution in Asia*. Boulder: Lynne Rienner.

de Janvry, A., M. Fafchamps, and E. Sadoulet. 1991. Peasant household behaviour with missing markets: Some paradoxes explained. *Economic Journal* 101: 1400–1417.

Haggblade, S., P.B.R. Hazell, and T. Reardon, eds. 2006. *Transforming the rural nonfarm economy.* Wallingford: CAB International.

Hayami, Y. 1996. The peasant in economic modernization. *American Journal of Agricultural Economics* 78: 1157–1167.

Hayami, Y., and M. Kikuchi. 2000. *A rice village saga: Three decades of green revolution in the Philippines*. London: Macmillan.

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

Johnson, D.G. 1951. Resource allocation under share contracts. *Journal of Political Economy* 58: 111–123.

Marshall, A. 1890. *Principles of economics*. 8th ed. London: Macmillan Press.

Otsuka, K. 2007. Efficiency and equity effects of land markets. In *Handbook of agricultural economics*, ed. R.E. Evenson, P.L. Pingali, and T.P. Schultz. Amsterdam: North-Holland.

Otsuka, K., and F. Place, eds. 2001. *Land tenure and natural resource management: A comparative study of agrarian communities in Asia and Africa.* Baltimore: Johns Hopkins University Press.

Quisumbing, A.R., J.P. Estudillo, and K. Otsuka. 2004. *Land and schooling: Transferring wealth across generation*. Baltimore: Johns Hopkins University Press.

Schultz, T.W. 1964. *Transforming traditional agriculture*. New Heaven: Yale University Press.

Singh, I., L. Squire, and J. Strauss, eds. 1986. *Agricultural household models.* Baltimore: Johns Hopkins University Press.

Pecuniary and Non-pecuniary Economies

J. de V. Graaff

The term *pecuniary* economies is probably due to Viner (1931), who distinguished them from *technological* ones. It is the latter that are often referred to as *non-pecuniary* economies. It is helpful to start by distinguishing internal economies from those that are external to the firm, and to deal with them separately.

Economies *internal* to the firm may be classified as pecuniary or non-pecuniary without much difficulty. Both (unless offset by diseconomies) give rise to falling unit costs as output is expanded. The former result from the firm's being able to negotiate lower prices for inputs bought in quantity; the latter from fuller utilization of 'lumpy' inputs such as buildings, machines and the services of specialists. In short, pecuniary internal economies have to do with factor prices; non-pecuniary ones with indivisibilities.

Economies *external* to the firm may also be classified into pecuniary and non-pecuniary (or technological), but there is scant connection with the previous definitions and some doubt as to the appropriateness of the nomenclature. In what follows we take the non-pecuniary ones first, as they give fewer difficulties; and use *economies* in a sense wide enough to include *diseconomies*.

Non-pecuniary external economies exist when there is technological interdependence between firms in the sense that it is not possible to specify one firm's production function without knowing at least one of the inputs or outputs of another firm. There are numerous wellworn examples in the literature, and many more in actual life. Among them are: (i) two wells on a single oilfield, the yield of the one depending on the rate at which the other is pumped; (ii) effluent from a pulp mill, increasing the purification costs of a brewer drawing downstream water; and (iii) a heavier crop of fruit in an orchard, resulting from better pollination by bees from a neighbour's apiary.

These examples illustrate some of the many sorts of technological inter-relationships that are possible between firms. All of them represent non-pecuniary external economies and (if, as in the examples, they operate at the margin) will produce divergences between marginal private and marginal social costs. These divergences give rise to certain problems in connection with the optimal working of the market mechanism that can be resolved by merger, by negotiation (if transactions costs are not too high) or by other measures such as taxes and subsidies.

Pecuniary external economies comprise the last of the four categories. They have been defined differently by different writers but are in essence what Marshall had in mind when he described external economies (he did *not* use the word) 'pecuniary') as 'those dependent on the general development of the industry' (1920, p. 266) or on 'advances made by subsidiary industries' (1920, p. 614). The examples that used to be given include the development of transport and telephone services, the growth of a skilled labour supply and such unlikely oddities as the appearance of a trade journal. Typically, they operate through time rather than in a framework of static equilibrium.

Viner (1931) made an attempt at formalizing the definition. For him pecuniary external economies exist when decreasing unit costs for a firm result from reductions in the prices paid by it for its factors of production when the industry of which it is a part grows as a whole. Most of these price reductions would presumably be due to internal economies in other firms; but Viner also has the rather fanciful illustration of labourers having a preference for working in an important rather than a minor industry and therefore being willing to accept a lower wage as their chosen one expands. Nevertheless, his definition gets close to Marshall's original idea.

Scitovsky (1954) makes his definition turn on profits rather than prices. Pecuniary external economies exist for him when a firm's profits depend 'not only on its own output and factor inputs but also on the output and factor inputs of other firms'. No treatment in terms of profits can exclude price changes, so the two definitions are not necessarily incompatible, although Scitovsky's is much wider. It is indeed so wide that it includes *non*-pecuniary external economies. It also includes all the aspects of mutual interdependence through the price system that are normally treated in General Equilibrium studies. There is very little that is left out. One is left wondering why one should be asked to call almost everything an external economy.

A possible reason is that the concept of pecuniary external economies has found its chief use in the field of economic development. There one is not dealing with equilibrium studies but with the growth of a whole economy over a broad front. It is difficult for private entrepreneurs to get sufficient information to coordinate their investment plans from current market prices. What they really would need would be a complete set of forward markets in 'future' goods through which the mutual interdependence of the various projects could find expression. In the absence of such markets there is a case for some form of central planning. The recognition of a need for planning is often regarded as an admission of market failure. But external economies of the technological kind are known to lead to market failure. So let us call both causes of failure by the same name: external economies.

That seems to be how the terminology has developed. Clarity would be served by reversing the process and speaking instead of technological interdependence (of production functions) on the one hand, and market interdependence (via the price system) on the other.

See Also

- ▶ Externalities
- ▶ Marshall, Alfred (1842–1924)
- ▶ Pigou, Arthur Cecil (1877–1959)
- ▶ Rising Supply Price
- ▶ Welfare Economics

Bibliography

Marshall, A. 1920. *Principles of economics*, 8th ed. London: Macmillan.

Scitovsky, T. 1954. Two concepts of external economies. *Journal of Political Economy* 62: 70–82.

Viner, J. 1931. Cost curves and supply curves. *Zeitschrift für Nationalökonomie* 3: 23–46.

Pecuniary Versus Non-pecuniary Penalties

Dan Kahan

Abstract

'Pecuniary' penalties (fines) seem underutilized relative to 'non-pecuniary' penalties such as imprisonment, since they are ceteris paribus cheaper for society to impose. But the public preference for imprisonment over fines might reflect the value that the public attaches to the condemnatory meaning that imprisonment, unlike fines, conveys. An economic theory of punishment should include this sensibility in the social welfare calculus used to appraise the efficiency of various forms of punishment. The expressive utility of imprisonment might more than offset the higher cost of imprisoning offenders who could just as effectively be deterred by fines.

Keywords

Becker, G.; Bentham, J.; Fines; Imprisonment; Pecuniary vs. non-pecuniary penalties; Posner, R.; Punishment; Social norms

JEL Classifications

C43; E51

The topic of 'pecuniary' and 'non-pecuniary' penalties involves a distinction easily grasped but also a puzzle not easily solved. The distinction is between fines and all other types of criminal punishments, most conspicuously imprisonment. The puzzle arises from the seeming underutilization of pecuniary penalties, especially relative to imprisonment, in the American criminal justice system.

An economic theory of law furnishes a straightforward case for the use of pecuniary penalties. From an economic point of view, it is assumed that an individual will refrain from criminality when the expected cost of lawbreaking exceeds the expected gains (Bentham 1843). The law can raise the expected cost by divesting offenders of either their liberty or their monetary assets. Depriving them of the latter, however, is much cheaper for society: whereas imprisonment demands an immense expenditure of resources, fining involves a transfer of wealth from offenders to the state. Accordingly, whenever a fine of a particular size and a prison term of a particular length would impose equivalent disutility on offenders, the state, in the interest of efficiency, should select the fine (Becker 1968).

But as cogent as it might be, this economic defence of pecuniary penalties has had strikingly little influence on the law. Non-violent offenders, who could be fined rather than imprisoned consistent with public safety, make up over half the American prison population. Many of these non-violent offenders are likely to be poor and thus effectively immune to the threat of massive fines (Shavell 1985). However, the possibility of 'day fines' – a procedure, common in Europe, whereby fines are meted out over time based on ability to pay – would make stiff penalties feasible even for offenders of relatively modest means. Based on these considerations, there is widespread expert consensus that American jurisdictions rely far too heavily on imprisonment relative to pecuniary penalties, particularly for white-collar offenders, who are obviously the least violent and the most credibly threatened with large fines (Morris and Tonry 1990; Posner 1980).

Confronted with this tension between theory and practice, one might be tempted to shrug one's shoulders at the seeming economic irrationality of the law and move on. But before doing so, it is worth considering whether the relative underutilization of pecuniary penalties might itself be explained in economic terms – ones that the conventional defence of pecuniary penalties overlooks.

Perhaps surprisingly, the key to a more complete economic analysis is rooted in a distinction that sociologists and philosophers draw based on the *social meanings* that legal impositions convey. 'Prices', on this account, refer to pecuniary exactions that connote an intention to levy a tax on an activity that society views as morally permissible; 'sanctions', in contrast, connote punishments that the state imposes on activities that are morally forbidden (Cooter 1984). Criminal fines, particularly for offences that seem to involve a serious flouting of societal norms, often strike members of the public, dissonantly, as mere 'prices'. Imprisonment, in contrast, unambiguously registers as a 'sanction'; by virtue of the veneration of individual liberty in American society, taking a person's liberty away conveys a highly condemnatory intent on the part of the law (Kahan 1996).

Is there any reason, economically speaking, to prefer sanctions to prices? Perhaps. Again, from an economic point of view, a person will refrain from criminality when the expected cost exceeds the expected gain. If that is correct, then the law can discourage criminality not just by increasing an offender's estimation of the costs but also by diminishing his or her valuation of the gains associated with lawbreaking. It is often argued, in fact, that the law plays a vital role in inculcating preferences that conduce to law-abiding behaviour (Andenaes 1966; Dau-Schmidt 1990).

On this account, one economic defence of imprisonment in preference to fines would be that sanctions are more effective than mere prices in instilling law-abiding preferences. Imposing a sanction, such as imprisonment, on an act would impart information – that the act is morally frowned upon – whereas a mere price, such as a fine, would not. On the assumption that individuals adapt their values to those expressed in law, the threat of imprisonment would in these circumstances more effectively suppress a potential offender's estimation of the gain associated with a particular criminal act than would the threat of a fine. If this characteristic of imprisonment is sufficiently pronounced, it might result in behavioural effects that more than compensate for the additional cost of imprisonment (Kahan 1997).

But such an argument is speculative. There is some empirical evidence that the perceived justice of legal outcomes and procedures influences persons' disposition to obey (Tyler 1990; Nadler 2005), but none to show that the form of punishment (abstracted from its severity) does.

In addition, the claim that the law prefers imprisonment to fines because of its superior preference-shaping effect has a 'just so' quality. Aside from the implicit and by now largely rejected assumption that the law tends naturally toward efficiency, the preference-shaping defence offers no explanation of how this supposed feature of imprisonment figures in the political economy of punishment selection. Accordingly, this argument does not offer a particularly satisfying solution to the puzzle of why American jurisdictions so decidedly favour pecuniary over non-pecuniary penalties.

The real contribution the 'price-sanction' distinction makes to solving this puzzle consists in its power to illuminate an otherwise obscure element of the public demand for punishment. Criminal punishments, that distinction reminds us, do more than protect society from harm; they also evince a societal attitude towards criminal wrongdoers. The preference for imprisonment over fines, then, might reflect the immense value that the public attaches to the condemnatory meaning that sanctions, relative to mere prices, express.

This hypothesis finds ample empirical support. Some of it is experimental: even when fines are perceived as imposing levels of disutility comparable to particular terms of incarceration, members of the public reject fines as lacking the power to express moral condemnation (Marinos 1997). Analysis of the reasoning of legislators, judges, and ordinary citizens confirms that it is this sensibility that causes legal decision-makers to resist substituting fines for imprisonment as a punishment for white-collar offences and for other serious but non-violent common crimes (Kahan 1996).

The public demand for expressively satisfying punishments arguably helps to acquit imprisonment of the charge that it is less efficient than fines. Members of the public clearly value criminal punishment not only as a device for protecting them from harm but also as a ceremonial gesture for proclaiming the deviant status of those who violate societal norms (Garfinkel 1956). There is no reason, economically speaking, to exclude this sensibility from the social welfare calculus used to appraise the efficiency of various forms of punishment. If the value that members of society obtain from the expressive utility of imprisonment is sufficiently high, that species of well-being might more than offset the higher cost of imprisoning offenders who could just as effectively be deterred by fines (Kahan 1998).

Even more importantly, the contribution that the 'prices-sanctions' distinction makes to solving the puzzle of non-pecuniary penalties suggests insights into how, from an economic perspective, the law might be profitably reformed. Once the full dimensions of the social welfare function of punishment is discerned, it becomes clear that making law more efficient requires identifying relatively cheap punishments that, unlike fines, are comparable to imprisonment in *both* their expressive *and* their deterrent value. Because the expressive inadequacy of fines is also what constrains their political acceptability, expressively adequate alternatives to imprisonment also stand a much better chance than do fines of being adopted in the political process. These arguments have been used to defend the advent of *shaming* punishments – another non-pecuniary penalty – for white-collar criminals and other common offenders (Kahan and Posner 1999).

The desirability of shaming or any other non-pecuniary penalty is obviously open to debate, economically and otherwise. What should not be, however, is the proposition that the perfection of economic analyses of law depends on their cognizance of the full range of societal benefits, including expressive ones, that the law secures.

See Also

- ▶ Becker, Gary S. (Born 1930)
- ▶ Deterrence (Theory), Economics of
- ▶ Law, Economic Analysis of

Bibliography

Andenaes, J. 1966. General preventive effects of punishment. *University of Pennsylvania Law Review* 114: 949–983.

Becker, G. 1968. Crime and punishment: An economic approach. *Journal of Political Economy* 76: 169–217.

Bentham, J. 1843. Principles of penal law. In *The works of Jeremy Bentham,* vol. 1, ed. J. Bowring. Edinburgh/London: W. Tait/Simpkin, Marshall & Co.

Cooter, R. 1984. Prices and sanctions. *Columbia Law Review* 84: 1523–1560.

Dau-Schmidt, K. 1990. An economic analysis of the criminal law as a preference shaping policy. *Duke Law Journal* 1990: 1–38.

Garfinkel, H. 1956. Conditions of successful degradation ceremonies. *American Journal of Sociology* 61: 420–424.

Gibbs, J. 1978. Preventive effects of capital punishment other than deterrence. *Criminal Law Bulletin* 14: 34–50.

Kahan, D. 1996. What do alternative sanctions mean? *University of Chicago Law Review* 63: 591–653.

Kahan, D. 1997. Social influence, social meaning, and deterrence. *Virginia Law Review* 83: 349–395.

Kahan, D. 1998. Social meaning and the economic analysis of crime. *Journal of Legal Studies* 27: 609–622.

Kahan, D., and E. Posner. 1999. Shaming white-collar criminals: A proposal for reform of the federal sentencing guidelines. *Journal of Law and Economics* 42: 365–391.

Marinos, V. 1997. Equivalency and interchangeability: The unexamined complexities of reforming the fine. *Canadian Journal of Criminology–Revue Canadienne De Criminologie* 39: 27–50.

Morris, N., and M. Tonry. 1990. *Between prison and probation: Intermediate punishments in a rational sentencing system*. New York: Oxford University Press.

Nadler, J. 2005. Flouting the law. *Texas Law Review* 83: 1399–1442.

Posner, R. 1980. Optimal sentences for white-collar criminals. *American Criminal Law Review* 17: 409–418.

Shavell, S. 1985. Criminal law and the optimal use of nonmonetary sanctions as a deterrent. *Columbia Law Review* 85: 1232–1262.

Tyler, T. 1990. *Why people obey the law*. New Haven: Yale University Press.

Pennington, James (1777–1862)

Murray Milgate

Keywords

Bank of England; Currency principle; Fullarton, John; Pennington, James; Tooke, Thomas

Pennington may be credited with having been among the first to produce a concise statement of the so-called currency principle which formed the basis of the thinking behind the Bank Charter Act of 1844. Pennington's proposal appeared in the form of a privately printed *Memorandum* issued in 1827. This tract actually contained two memoranda, separated by a reply to the first (of 1826) from Huskisson. Much of the material from the memoranda was subsequently reissued by Pennington himself in 1840 as part of his larger *Letter to Kirkman Finlay, Esq., on the Importation of Foreign Corn*. It seems likely that the first memorandum was written at the suggestion of Thomas Tooke.

Pennington's argument was that by bringing under the direct control of the Bank of England the entire note issue, and by restricting that issue to the amount of specie reserves of the central bank so that, as Pennington put it, 'in all cases paper would contract and expand according to the increase or diminution of its bullion', monetary stability would be ensured. The similarity between this proposition and the practices which were emerging in the Bank of England itself at roughly the same time is worth noting. The so-called 'Palmer rule' differed from Pennington's only in as much as that under its operation the monetary magnitude that was to be tied to the Bank's specie reserves included not only notes and coin but also deposits. However, unlike the Palmer rule, Pennington's proposal entailed control by the central bank over the independent note-issuing activities of the country banks.

Though Pennington's proposal did not gain much public notoriety at the time, by the early 1830s Pennington had become an occasional adviser to ministers of state and to government departments. By 1844, it would seem that Pennington was sufficiently close to the government to have been asked to assist in drafting the technical details of the Bank Charter Act. The evidence currently available, however, suggests that this assistance was requested after Peel had decided upon the main provisions of that Act.

It must be remembered that although an advocate of the currency principle, Pennington actually opposed the division of the Bank of England into separate Issue and Banking departments (as was done under the provisions of the Act of 1844). A note to this effect written by Pennington was appended by Tooke to the first volume of his *History of Prices*. The curious fact that the most famous opponent of the currency principle should have appended to his celebrated study a note by the originator of that principle, was used to humorous effect by some of Tooke's adversaries (in particular, Torrens) in subsequent controversy over the consistency of Tooke's arguments. There is a comment on this aspect of the debate by Fullarton in his *Regulation of Currencies* (1844; 2nd edn 1845, p. 191).

Pennington was born at Kendal on 23 February 1777, and died at Clapham Common on 23 March 1862. There is an admirable and thorough survey of Pennington's life and work written by R.S. Sayers to accompany his edition of Pennington's economic writings, in which can be found all of the tracts referred to above. It may be of anecdotal interest to record that Hayek has conjectured that Pennington's brother may have been the apothecary who attended Henry Thornton during his final illness in 1815 (1939, p. 33n).

Selected Works

1829. Deposits with bankers. In *A letter to Lord Grenville . . . on the value of the currency,* ed. T. Tooke. London: J. Murray, 1829.

1838–57. Appendix C. In *A history of prices,* ed. T. Tooke, vol. 2. London: Longman, Orme, Brown, Green & Longmans.

1840. *A letter to Kirkman Finlay, esq., on the importation of foreign corn ….* London: Simpkin, Marshall & Co.

1848. *The currency of the British colonies*. London: W. Clowes & Sons.

1963. In *The economic writings of James Pennington*, ed. R.S. Sayers. London: London School of Economics and Political Science.

Bibliography

Fullarton, J. 1844. *On the regulation of currencies*. London: John Murray.

Hayek, F.A. 1939. Introduction to Henry Thornton, *An enquiry into the nature and effects of the paper credit of Great Britain* (1802). London: George Allen & Unwin. Reprinted New York: Kelley, 1962.

Tooke, T. 1840. *A history of prices and of the state of the circulation in 1838 and 1839, with remarks on the corn laws and some of the alterations in our banking system*. London: Longman, Orme, Brown, Green & Longmans.

Penrose, Edith Tilton (1914–1996)

R. L. Marris

Keywords

Firm, growth of the; Firm, theory of the; Managerial capitalism; Penrose effect; Penrose, E. T.

JEL Classifications

B31

Edith Penrose had a distinguished career in economics teaching, research and administration in the USA and the UK. From Johns Hopkins University she went to London University, where she became a professor; later she was Associate Dean of Research and Development at INSEAD in France. In administration she also rendered valuable service to the UK as chairman of the economics committee of the British Social Sciences Research Council. She retired in the mid-1980s.

In research, in the final part of her career, she concentrated on the oil industry and on multinational companies generally. Her place in the history of economic thought, however, lies in a single book *The Theory of the Growth of the Firm,* published in 1959. The review in *The Economic Journal* (1961) predicted that the book would prove one of the most influential books of the decade: this proved an understatement.

In Edith Penrose's conception, a firm is an administrative organization representing a collection of human and material resources for the purpose of producing goods and services for sale on the markets. It is essentially directed and controlled by its managers who will for various reasons be strongly motivated towards growth. The firm is not confined to any one product or market, but may diversify as its managers think fit. Profits, as seen by Penrose in her original book, were essentially a means to that end, a necessary condition for expansion.

There were, however, important administrative restraints on the rate of growth. Human resources required for *the management of change* (*growth*) were firm-specific and therefore, at any one moment, internally scarce. Expansion, however, included recruitment of additional high level human resources, that is, recruitment of additional growth-creating capacity. Therefore, subject to the dynamic constraint, there need be no ultimate limit on size. More generally, the relationship can be stated as a proposition that the *level* of current efficiency will, beyond a point, diminish with the rate of change of size: fast growth has a price.

The book was also rich in many associated and diverse ideas which cannot be set out in detail. The administrative 'Penrose effect' has been generally accepted and incorporated into a variety of micro-and macroeconomics, especially in the field known as 'the Corporate Economy'. The idea was most especially used in Robin Marris in *The Economic Theory of Managerial Capitalism* (1964) and by Hirofumi Uzawa in a significant contribution to macroeconomics a few years later (Uzawa 1969).

The total effect of Edith Penrose's work was that of destruction of the neoclassical model of the firm, followed by reconstruction. In the following years, however, despite the wide recognition the work received, classroom microeconomic theory, and also classroom industrial organization, often seemed to continue as if nothing had happened.

Selected Works

1959. *The theory of the growth of the firm.* Oxford: Basil Blackwell.

Bibliography

Marris, R.L. 1964. The economic theory of managerial capitalism. London: Macmillan; New York: Free Press.
Uzawa, H. 1969. Time preference and the Penrose effect in a two-class model of economic growth. *Journal of Political Economy* 77: 628–652.

Pension Systems: Principles, Debates and Analytical Errors

Nicholas Barr

Abstract

Many countries face problems financing pensions in the face of population aging. There is controversy about the underlying economic theory, the extent of the problem and the best mix of policies to protect old-age security. This article sets out the economic analytics of pensions, discussing in turn their multidimensional nature, principles of analysis, the reasons why government gets involved, and debates and analytical errors. Those analytical errors matter, because they lead to policy errors. A central conclusion is that although there are sound principles of pension design, there is no single best pension system for all countries.

Keywords

Consumption smoothing; Defined-benefit; Defined-contribution; Funding; Insurance; Notional defined-contribution; Pay-as-you-go; Pensions; Risk sharing

JEL Classification

H10; H30; H31; H55

This article discusses the multidimensional nature of pensions, principles of analysis, the reasons why governments get involved, and debates and analytical errors.

A Multidimensional Set of Issues

Analysis of pension systems has to accommodate different ways of organising pensions, multiple objectives, multiple risks, different ways of relating contributions and benefits, and different ways of adjusting contributions and benefits over time.

Multiple Ways of Organising Pensions

Pensions can be fully funded, pay-as-you-go (PAYG), or partially funded.

Fully-funded pensions are paid from an accumulated fund built up over a period of years out of contributions by and/or for members. Funding is thus a method of accumulating financial assets, which are exchanged for goods at some later date.

PAYG pensions are paid out of current contributions. They are usually run by the state, on the basis that the state can, but does not have to, accumulate assets in anticipation of future pension claims, and can tax the working population to pay the pensions of the retired generation. From an economic viewpoint, a publicly organised PAYG system can be looked at in several ways. As an individual contributor, a worker's claim to a pension is based on legislation that, if she pays contributions now, she will receive a pension in the future. From an aggregate viewpoint, the state is taxing one group of individuals and transferring the revenues to another and, in that sense, the arrangement is little different from other income transfers.

Pensions can also be partially funded, either as part of their long-run design or because they can accommodate less-than-full funding to spread risk during times of adjustment.

Multiple Objectives

When retirement provisions were introduced in the nineteenth century, someone aged 65 was typically infirm and interfered with the productivity of younger workers. The purpose of pensions was to clear out unproductive older workers, so it made sense for retirement to be mandatory and complete. Over time, people have been living longer and countries have grown richer, making it possible to give people a period of leisure at the end of their working lives. Thus the purpose of retirement has changed: it is no longer simply a device for getting rid of dead wood, but a social construct for dividing the life cycle into working life and leisure. For this latter purpose, it is right to recognise that individuals vary widely in their preferences and circumstances. Many people do not want to retire fully as soon as they are allowed, because of the extra earnings, because of extra pension and/or because they continue to enjoy working. Pension design needs to recognise these new complexities.

From the viewpoint of individuals and families, income security in old age requires two sets of instruments: a mechanism for smoothing consumption, and a means of insurance:

- **Consumption smoothing:** a central purpose of retirement pensions is to enable a person to transfer consumption from her earnings in middle years to her retired years, allowing her to choose a better time path of consumption over working and retired life.
- **Insurance:** pensions can provide at least partial insurance against a range of risks. Annuities address the longevity risk. Individuals also face risks to future earnings during working life. These risks can be insured in part through unemployment and disability insurance, but they also have consequences for retirement, which pension systems can address at least partly through a redistributive element.

Public policy has objectives additional to improving consumption smoothing and insurance.

- **Poverty relief:** pension systems target resources on people who are poor on a lifetime basis, and are thus unable to save enough. As a practical matter, poverty relief also has to address transient poverty, either through programmes specifically for the elderly or through a wider programme for poverty relief. In some ways, the design of poverty relief for older people is simpler in that transfers to a group that has partly or wholly stopped paid work are less likely to weaken work incentives.
- **Redistribution:** lifetime redistribution can be achieved by paying pensions to low earners that are a higher percentage of their previous earnings. Since lifelong earnings are uncertain from the perspective of an individual, such a system provides some insurance against low earnings. There can also be redistribution towards families: for example paying a higher pension to a married couple than to a single person.

Pension systems can also redistribute across generations. This element has been common in the startup of PAYG pension systems, because people who worked before the system was created would otherwise have had no or low retirement incomes.

Alongside these primary objectives, pensions may have secondary goals, including economic growth or, less stringently, avoiding undercutting economic growth. There is debate about the relative weights accorded to old age security and to these secondary objectives.

Multiple Risks

It is helpful to distinguish different elements.

- Individual (or idiosyncratic) risk concerns the distribution of a given average risk across individuals, for example the risk that an airline will lose a person’s bags.
- Systemic risk (or common shocks) arise when the average risk changes; such risks affect all or

many individuals. Inflation, for example, affects everyone.

Some risks have both elements: a person aged 65 faces a probability distribution of remaining life expectancy (individual risk), but average remaining life expectancy can rise over time (systemic risk).

The risks facing an individual can loosely be divided into systemic risks, market risks and risks connected with individual behaviour.

- **Systemic risks.** Macroeconomic risk affects output, prices or both. Demographic risk arises through longer life expectancy and lower fertility. Political risks can arise even in well-governed countries.
- **Market risks** arise from systemic shocks, but also have idiosyncratic elements:
  - Earnings risk: a worker's earnings profile has deterministic elements (e.g. the decision to invest in human capital) and stochastic elements, relating to labour markets and health risks.
  - Investment risk: accumulations held in the stock market are vulnerable to market fluctuations. At its extreme, if a person with a fully funded individual account is obliged to retire on her 65th birthday, there is a lottery element in the value of her pension accumulation.
  - Annuities market risk: for a given accumulation, a person's annuity at a given age will be affected by the life expectancy of his birth cohort and the discount rate used by the annuity provider. Annuity providers can also fail.
- **Risks connected with individual behaviour**
- Principal risk arises through bad decisions by participants. As discussed below, poor choices can arise from imperfect information, and also for reasons which behavioural economics explains.
- Agency risk can arise through incompetent or fraudulent fund management. More importantly, managers in private systems may have different incentives from plan participants.

Many of these elements face policy makers not only with risk (where the probability distribution of outcomes can be estimated fairly precisely), but also with uncertainty, where the probability distribution of outcomes is not well known. Actuarial insurance can address risk, but faces problems with uncertainty.

Multiple Ways of Relating Contributions and Benefits

Whether funded or PAYG, a separate question is how closely pension benefits are related to a worker's previous contributions. Three approaches are common.

- **Defined-contribution (DC) plans.** In a pure defined-contribution system (i.e. one with no redistribution across individual accumulations), a person's consumption in retirement, given life expectancy and the rate of interest, is determined by the size of his or her lifetime pension accumulation.
- **Defined-benefit plans.** In a defined-benefit (DB) plan, a worker's pension is based not on his or her accumulation, but on the worker's wage history. The sponsor's contribution is conceptually the endogenous variable ensuring the system's financial balance.

Defined-benefit systems can be structured in different ways. A key design feature is the way in which wages enter the benefit formula. In a final salary system, pensions are based on a person's wage in his or her final year or final few years. Alternatively, the pension can be based on a person's wages over an extended period, including a whole career.

A second design feature is the rules which specify how the level of benefits changes when a worker delays claiming a pension. Such adjustments may or may not be actuarial. Third, defined-benefit systems can be run by the state or by employers.

- **Notional defined-contribution (NDC) plans.** This arrangement is conceptually similar to defined-contribution plans in that contributions are notionally 'accumulated' to determine a balance which is converted into an annuity at retirement, but different in that that they are not fully funded and may be almost entirely PAYG.

NDC plans parallel defined-contribution plans:

- A workers pays a contribution of *x*% of his or her earnings, which is credited to a notional individual account, that is, the state 'pretends' that there is an accumulation of financial assets, though in reality the account balance is for record keeping only.
- The cumulative contents of the account are credited with a notional interest rate, specified by legislation.
- At retirement, a person's notional accumulation is converted into an annuity, such that the present value of benefits (given the worker's age and the remaining life expectancy of his or her birth cohort) is equal to the value of the person's notional accumulation, using the notional interest rate as the discount rate.

Multiple Ways of Adjusting Contributions and Benefits Over Time

In the face of the risks outlined above, any pension system must adapt to actual developments over the medium term. Adjustment can be on the contributions side, on the benefit side, or both.

- **Increasing the income of the pension system.** Pensions can adjust through:
  - Higher savings by current workers (fully funded individual accounts) or higher contributions by today's workers (a less-than-fully funded definedbenefit system). Thus the extra revenue comes from today's working participants.
  - Higher contributions by the plan sponsor (a defined-benefit plan provided by a firm or industry). Depending on elasticities in labour, capital and product markets, the extra revenue can come from current workers (through effects on wage rates), shareholders and the taxpayer (through effects on profits), customers (through effects on prices) and/or past or future workers, if the company uses surpluses from some periods to boost pensions in others.
  - Higher contributions by insurance companies where retirees or plan sponsors have bought annuities. In that case, again depending on the relevant elasticities, the extra revenue comes from the insurance company's workers, shareholders or customers.
  - Higher contributions by today's taxpayers (a public pension). Thus the extra income comes from today's taxpayers and hence, through government borrowing, can also come from future taxpayers, allowing intergenerational risk sharing.
- **Reducing pension spending.** Total pension spending is the product of (a) the level of the average pension and (b) the number of pensioners. A major determinant of the latter is the earliest age at which a person can begin to draw pension. Policies to reduce pension spending can operate on either or both.
  - The monthly pension at a given eligibility age can be reduced in several ways. A lower rate of accrual during working life can be implemented through (a) a lower return to financial assets (a fully funded definedcontribution pension), or (b) a less generous legislated accrual rule. Lower pensions in payment can be implemented through less generous indexation of pensions in payment, or a reduction in pension.
  - An increase in earliest eligibility age affects workers but not retirees. With less-than-actuarial adjustment, total spending on pensions declines, e.g. a defined-benefit plan. With actuarial adjustment, there is no saving in total pension spending, but a given volume of spending can maintain a desired replacement rate (fully funded defined-contribution pensions or an NDC system). In this case, the purpose of an increase in eligibility age is to address adequacy rather than sustainability. In either case, rules are needed about how benefits increase where someone chooses to start benefit later than the earliest eligibility age.

Principles of Analysis

No Single Best Pension System

Pension systems have the multiple objectives noted earlier. The pursuit of those objectives faces a series of constraints, including fiscal capacity, institutional capacity and the empirical value of behavioural parameters, such as the responsiveness of labour supply to the design of the pension system, and the effect of pensions on private saving. A further constraint is the shape of the pre-transfer income distribution: a heavier lower tail increases the need for poverty relief. Countries also face political and historical constraints.

There is no single best system for all countries, because policy makers at different times and in different places will attach different relative weights to the different objectives and the pattern of constraints will differ across countries. If objectives differ and constraints differ, the optimum will generally differ.

Though there is no single best system, there are clear principles of analysis.

- **A holistic view.** Analysis should consider the pension system as a whole. Pensions have a wide range of effects, including on the labour market, saving, economic growth, the distribution of risk and the distribution of income, including effects by gender and generation. What is relevant for analysis is the combined effect of the system as a whole. Thus it is necessary to consider together the parts of the pension system that provide poverty relief and those where the primary focus is the pursuit of other objectives.
- **Second-best analysis.** Simple theory assumes that individuals make optimal choices and that labour markets, savings institutions and insurance markets function ideally. Those assumptions do not apply to pension systems. As discussed below, workers and pensioners face information problems, behavioural problems and missing markets, as well as factors broader than pensions, such as the inescapable existence of distortionary taxation.

  Framing the argument in second-best terms starts from the multiple objectives of pension systems. Thus policy has to optimise (not minimise or maximise) across a range of objectives, which cannot all be achieved fully at the same time. Policy has to seek the best balance between consumption smoothing, poverty relief and insurance, a balance that will depend in each society on the weights given to those and other objectives and to the different constraints that societies face.
- **Not just economics.** The optimal pension design depends not only on good economics, but also on good politics. The political economy question is how to reconcile the long-run objectives of pensions with short-run economic and political pressures. Different designs give governments more or less scope for adjusting the system; greater flexibility allows wider risk sharing, but can face the risk of government failure. Thus the choice of pension design will depend in part on the weight that policy makers give to wider risk sharing and in part on an empirical view of the quality of government in the country concerned.

  Alongside discussion of what design might be optimal is the equally important question of what is feasible. The greater a country's fiscal and institutional capacity the wider the range of feasible pension designs. Many reforms have come to grief not because of conceptual flaws, but because reformers were over-optimistic about economic circumstances or the ability to administer the new system effectively.

Why Does Government Get Involved?

To address Information and Behavioural Problems

The economics of information explains why the model of the well-informed consumer does not hold in many areas of social policy. There is considerable evidence of poor information about pensions. A survey revealed that 50% of Americans did not know the difference between a stock and a bond. Most people with an individual account do not understand the need to shift from equities to bonds as they age. And few people realise the significance of administrative charges for pensions. A system which offers wide choice is administratively costly: with an individual account, over a full career an annual management charge of 1% of the individual's accumulation reduces the accumulation (and hence the pension) by 20% (Barr and Diamond 2008, Box 9.4).

Recent lessons from behavioural economics also yield powerful lessons, explaining such phenomena as procrastination (people delay saving, do not save, or do not save enough), inertia (people stay where they are) and immobilisation (where conflicts and confusion lead people to behave passively, like a rabbit in a car headlight).

These bodies of theory suggest that in the context of pensions the benefits of wide choice are limited and, for most people likely to be outweighed by the costs of choice. These considerations suggest a series of guidelines for the design of individual accounts:

- Use automatic enrolment.
- Keep choices simple: for most people, highly constrained choice is a deliberate and welfare-enhancing feature of good pension design.
- Design a good default option for people who make no choice.
- Decouple account administration from fund management, with account administration centralised and fund management organised on a wholesale, competitive basis.

The US Thrift Savings Plan for federal civil servants (www.tsp.gov) complies with these criteria. The plan offers participants a limited choice of portfolios. In 2007 workers could choose from six funds, including a life-cycle option (i.e. an option in which a person's portfolio shifts automatically from mainly equities to mainly bonds as he or she ages). A government agency keeps centralised records. Fund management is on a wholesale basis. Investment in private sector assets is handled by private financial firms, which bid for the opportunity, and which manage the same portfolios in the voluntary private market, providing some insulation from political interference.

The plan (a) simplifies choice for workers, includes (b) automatic enrolment and (c) a default option, and (d) keeps administrative costs low, thus respecting information and behavioural constraints. The system of NEST pensions (www.nestpensions.org.uk) being introduced in the UK is a similar arrangement.

To address Missing Markets

A second role of government is to address missing markets. Two areas are particularly salient.

- **Indexed government bonds.** Efficient consumption smoothing allows workers to plan their future consumption. Thus workers need an instrument to provide a given real income in retirement. To achieve that, workers need to be able to protect themselves against inflation, particularly once the pension is in payment. However, inflation is a common shock, so that the private sector has problems providing risk-free protection against inflation. An important role for government is to provide indexed bonds that pensioners or pension funds can buy.
- **Annuities.** Both public and private projections tend consistently to underestimate increases in life expectancy. As a result, annuity providers make losses and either leave the market or price future annuities cautiously, giving pensioners poor value for money. One way to address the problem is for government to sell longevity bonds, allowing annuity providers to cover the risk that the average life expectancy of a cohort will exceed that which is predicted. In this arrangement, in (say) 2020, an insurance company would sell an annuity to an individual aged (say) 70 priced on official estimates of the remaining life expectancy of a 70-year old person in 2020, and insures against the cohort living longer than the 2020 projection by buying longevity bonds. If the cohort of annuitants lives longer than the 2020 projection the taxpayer finances the resulting extra cost through the longevity bonds. Thus the insurance company takes on the risk, the

taxpayer the uncertainty. This is a sensible division of labour. The role of government is to fill the missing market.

To Provide Poverty Relief

Though consumption smoothing and some forms of insurance can be provided by the private sector (e.g. fully funded individual accounts together with annuities), a pension system also needs to provide poverty relief. That element will require government involvement even in a system which is otherwise mainly private.

To Widen Options for Risk Sharing

Different pension designs and different forms of adjustment have different implications for the pattern of risk bearing.

In a pure fully funded defined-contribution system there is no redistribution across generations. Thus a cohort is constrained by its own past savings so that, in present-value terms, a representative individual gets out of a funded system no more than she has put in. Though annuities protect the individual against the risks associated with longevity, a pure defined-contribution system leaves her facing a wide range of risks associated with varying real rates of return to pension assets, the risks of future earnings trajectories and the future pricing of annuities. Thus risk falls entirely on the individual worker's future pension. It is possible to share risks more widely by adjusting current as well as future benefits, but in a fully funded system such risk sharing is only among current participants.

In a defined-benefit state system financed entirely from contributions, the risk of adverse outcomes falls on contributions, i.e. on current workers, and none of the risk directly on pensioners.

In a defined-benefit state system in which contributions can be supplemented by taxpayer support, the risk of adverse outcomes can be shared widely. The risk can fall on pensioners (e.g. by less generous indexation); on workers, through higher contributions; on current taxpayers; or (through government drawdown of past surpluses or increased borrowing) on past, present and/or future taxpayers.

In a pure defined-benefit system provided by a firm or industry, the risk of varying rates of return to pension assets falls on the employer, and hence, as discussed, on some combination of the industry's current workers, shareholders, customers and/or its past or future workers.

A major implication of any less-than-fully funded system, is that it relaxes the constraint that the benefits received by any generation must be matched by its own contributions. Thus, in sharp contrast with fully funded arrangements, a system with a PAYG element can redistribute across generations and can share risks across generations.

Debates and Analytical Errors

Debates

There is considerable controversy over the relative merits of PAYG and funded systems. There are debates about the right economic model; empirical magnitudes; the extent of a country's institutional capacity; the political economy of reform (for example, whether citizens regard their pension as safer based on a promise by government or as the owners of accounts); and ideology (for example about the role of the state).

The World Bank was a powerful advocate of funded systems in the 1990s and early 2000s, growing out of its influential publication *Averting the Old Age Crisis* (World Bank 1994). The book's many critics argue that its strength is in its diagnosis of the problems facing many pension systems, but its central weakness is that the book's analysis did not substantiate its central prescription – mandatory individual funded accounts. Although the debate continues, it can be argued that individual funded accounts have not stood the test of time, exemplified by countries which introduced individual funded accounts in the 1990s but did not follow through (China), or liquidated them (Argentina in the early 2000s, Hungary after the financial crisis of 2007) or partially liquidated them (Poland).

Proponents of individual funded account point to Chile, which has had the system since 1981. Barr and Diamond (2008) argue that Chile is a special case. It introduced funding at a time of budgetary surplus and over the years has displayed greater institutional capacity than countries at a similar level of development. The fact that Chile sustained the system does not mean that other countries will be able to do so. In addition, Chile came to recognise that individual accounts are not a complete pension system, but only part of one (the understated element being poverty relief) and set up a Presidential Commission (the Marcel Commission) which recommended a system of non-contributory pensions (introduced in 2008) to complement individual funded accounts. A further Presidential Commission was appointed in 2014 to consider reform of the system of individual accounts.

Analytical Errors

Discussion of pensions is prone to analytical errors.

Tunnel Vision

Analysis that focuses, often implicitly, on a single objective such as consumption smoothing may be flawed because it pays inadequate attention to other objectives such as poverty relief and gender balance. Similarly, it is generally mistaken to consider one part of the pension system in isolation, ignoring the effects of other parts. There is no efficiency gain from moving redistribution from one part of the system to another, even if the change leaves one part with no deviation from full actuarial principles.

Improper Use of First-Best Analysis

It is a mistake to focus on the labour market distortions caused by a given set of pension arrangements while ignoring or downplaying the contributions of those arrangements to the various goals of pension systems – contributions that are not available without distortions. A pension system that includes poverty relief will be distorting; minimising distortions implies minimising poverty relief – the cure is worse than the disease. Of course, pensions should be designed to avoid larger distortions than are justified by their contribution to goals, but minimising distortions is not the right objective.

Improper Use of Steady State Analysis

It is mistaken to focus on the design of a reformed pension system in a steady state while ignoring or underplaying the steps that are necessary to get to that steady state. This issue becomes especially important when considering whether or not to move from PAYG toward funded pensions. For example, US Social Security is less than fully funded because earlier cohorts were paid higher benefits than their contributions could have financed. The purpose of paying higher benefits was to raise the consumption of those cohorts after retirement.

If one does not include the value to earlier generations of this extra consumption, the analysis implicitly makes a long-run comparison; that is, it compares the economic situation in the USA today with what it would have been in an alternative long run with funded pensions. Thus the underlying question is: how does welfare in long-run state B differ from that in long-run state A? For a policy choice the appropriate analysis asks a different question: what are the welfare effects of moving from state A to state B? Either question is coherent. What is not legitimate is to take the answer to one question and apply it to the other.

Incomplete Analysis of Implicit Pension Debt

Some commentators point to the present value of all future promised pensions (a large figure) as being necessarily a problem. The analytical flaw in that approach is that it looks only at future liabilities (i.e. future pension payments) while ignoring explicit assets and the implicit asset of the government's ability to levy taxes. Too narrow a focus on costs also ignores the considerable improvement in people's wellbeing from increased old-age security. Just as public debt never needs to be fully paid off so long as the debt-to-GDP ratio does not get too large, so publicly provided pensions need not be fully funded, as long as the unfunded obligations are not excessive relative to the contributions base.

Incomplete Analysis of the Effects of Funding

This error appears in a number of guises. One example is to focus analysis purely on financial assets, ignoring the fact that what matters to pensioners is consumption. With the exception of housing, a pensioner's consumption in old age will depend on the output of goods and services produced by younger workers. PAYG and funding are both ways of organising claims on that output. It is therefore mistaken to focus excessively on how pensions are financed while ignoring future national output and its division between workers and pensioners. The error in this focus is its failure to recognise that the effects of funding on future output will depend on the answers to a series of questions, many of which are often addressed incompletely or ignored:

- Will funding pensions increase saving?
- Is increased saving the right objective?
- Will the regulatory changes that may (and generally should) accompany funding strengthen the performance of capital markets?
- If so, is it necessary for this purpose that pensions are mandatory?
- Are redistributive effects across generations from a move towards funding, discussed below, desirable policy?

Ignoring Distributional Effects

Because pension systems can redistribute across cohorts, it is necessary to consider who gains and who loses. It is a major error to ignore the fact that any choice between funding and PAYG *necessarily* makes choices about redistribution across generations. The point is most obvious if policymakers are establishing a brand new pension system. If they introduce a PAYG system, the first generation of retirees receives a pension, but returns to subsequent generations are lower; if they introduce a fully funded system, later generations benefit from higher returns, but the first generation receives little or no pension. Thus it is mistaken to present the gain to pensioners in later generations as a pure welfare gain, since it comes at the expense of the first generation. The same argument applies in a country that already has a PAYG system: a policy to move toward funding through higher contributions or lower benefits redistributes from current generations to future ones.

Whatever the merits of a move towards funding, the error in ignoring distributional effects is profound: it leads to mistaken claims for the superiority of some policies, and ignores the fact that a PAYG element in a pension system is generally welfare-enhancing because of the resulting possibility of intergenerational risk sharing.

These analytical errors matter: analytical errors lead to policy errors, many of which are identified in a World Bank evaluation of its own pensions work (World Bank 2006).

Disclaimer

The article draws heavily on writing with Peter Diamond (Barr and Diamond 2008, 2009, 2010). The responsibility for the views expressed and remaining errors is entirely mine.

See Also

▶ Individual Retirement Accounts
▶ Pensions
▶ Population Ageing
▶ Retirement

Bibliography

Barr, N., and P. Diamond. 2008. *Reforming pensions: Principles and policy choices*. New York: Oxford University Press.

Barr, N., and P. Diamond. 2009. Reforming pensions: Principles, analytical errors and policy directions. *International Social Security Review*, 62(2): 5–29 (also in French, German and Spanish).

Barr, N., and P. Diamond. 2010. *Pension reform: A short guide*. New York: Oxford University Press.
World Bank. 1994. *Averting the old age crisis*. Washington DC: World Bank.
World Bank. 2006. *Pension reform and the development of pension systems: An evaluation of World Bank assistance*. Washington, DC: Independent Evaluation Group, World Bank. http://lnweb90.worldbank.org/oed/oeddoclib.nsf/DocUNIDViewForJavaSearch/43B436DFBB2723D085257108005F6309/$file/pensions_evaluation.pdf.

Pensions

Zvi Bodie

Abstract

Pensions are benefit contracts that replace a person's earnings after she reaches old age and retires from the labour force. Pension systems vary widely across countries, but everywhere the government's role is to provide a minimum through a mix of cash and medical benefits. Governments often provide tax incentives for employers and unions to sponsor occupational pension plans that complement the government-run system. The nature of the pension benefits promised and the assets that back them have profound effects on social welfare, on the development of a country's domestic asset markets, and on the global financial system.

Keywords

Adverse selection; Annuities; Defined-benefit pension plans; Defined-contribution pension plans; Equipment trusts; Free-rider problem; Funding of pensions; Informational inefficiencies; Life-cycle saving model; Longevity insurance; Moral hazard; Occupational pensions; Pay-as-you-go pension plans; Pensions; Retirement; Risk sharing; Social Security (USA)

JEL Classifications

G23

Pensions are retirement income contracts, and their manifest function is to replace a person's earnings after she reaches old age and retires from the labour force. Prior to the Industrial Revolution, the extended family was the primary institution that performed this function. Elderly family members lived and worked with offspring on a family-owned farm, and all drew a common livelihood from it. In many of today's less developed countries, this family-based pattern for old-age support still holds true.

Over time, urbanization and other fundamental economic and social changes gave rise to new institutional structures for the care and support of the elderly in much of the industrialized world. An often-used metaphor for describing developed countries' pension systems is that of the 'three-legged stool'. The first leg consists of government-provided old-age assistance and insurance programmes; the second leg is comprised of employer or labour union-provided pensions; and the third is individual and family support. There is substantial variation in the mix of the three sources of retirement income, both across households in a given country and across different countries (Bodie and Davis 2000).

Pensions should be analysed in the context of a life-cycle model of saving. In this framework, people save during their working years so that they can consume in their non-working retirement period. Some simplifying assumptions can quickly convey the essence of the life-cycle approach. Assume for the sake of illustration that an individual enters the labour force at age 20, works until retiring at age 65, and dies at age 80. His initial wealth is zero. During the working years, he earns constant real labour earnings, a portion of which is saved for retirement. The saving includes personal saving and the accrual of benefits under social security and employer-sponsored pension plans. We assume that the individual chooses to save an amount during the working years sufficient to make his level of real consumption after retirement equal to what it was before retirement. These savings earn a zero real rate of interest. At retirement, a constant real retirement benefit is paid, and at death there is nothing left over as a bequest.

These assumptions imply that the ratio of consumption to earnings must equal the ratio of years of work to total years of work and retirement:

$$\begin{aligned}&\text{Years of work} \times (\text{earnings} - \text{consumption})\\&= \text{years of retirement}\\&\times \text{consumption Years of work} \times \text{earnings}\\&= (\text{years of work} + \text{years of retirement})\\&\times \text{consumption}\end{aligned}$$

$$\frac{\text{Consumption}}{\text{Earnings}} = \frac{\text{years of work}}{\text{years of work} + \text{years of retirement}}$$

In this example, there are 45 years of work and 15 years of retirement, so the ratio of consumption to earnings is equal to 45/60 or 75 per cent, and the individual's 'gross saving' rate during his working years is 25 per cent. The benefits received during retirement come from three sources corresponding to the components of gross saving: social security, employer-provided pensions, and personal saving.

The Government's Role in Providing Retirement Income

The government's role in providing retirement income varies considerably across countries, but despite these variations there is a common theme: in virtually every country the government provides a 'floor' of income protection for the elderly, with the aged population's needs met by some mix of national insurance and national welfare systems, in the form of cash and medical insurance. This floor (or 'safety net') is usually mandatory and cannot be transferred.

Several economic arguments justify the government's provision of a layer of retirement benefits for everyone (Merton 1983). The first deals with *informational inefficiencies*. It is costly to acquire the knowledge necessary to prepare and carry out long-run plans for income provision. Although peoples' lifetime financial plans depend on their individual preferences and opportunities, their goals may be similar enough that a standard retirement savings plan can prove suitable to many. By providing a basic plan that supplies at least a minimum level of old-age support, the government is likely to help people save more efficiently than they could on their own.

The second argument revolves around *adverse selection* problems, There is considerable 'longevity risk' that people will outlive their retirement savings because their date of death is not known with certainty, in contrast to the simplified version of the life-cycle model we described earlier. One way to insure against the risk of exhausting one's savings during retirement is to purchase a life annuity contract. But the private market for life annuities suffers from adverse selection because people with a higher-than-average life expectancy have a high demand for this kind of insurance. As a consequence, an average individual will find the equilibrium price for privately purchased life annuities too high, and will tend to self-insure against longevity risk by having an extra reserve of retirement savings. Universal and mandatory social security is one way of overcoming this adverse selection problem. Making participation in the national plan mandatory and not giving anyone a choice about the form of benefit payouts creates more complete pooling of longevity risk.

A third reason for a government-mandated universal retirement income system is to address the *free-rider* problem, which arises when the citizenry collectively feels an obligation to offer a universal 'safety net'. If this collective commitment were well understood by all, some people would avoid saving for their own retirement, intending instead to rely on benefits provided by others when they are old. Similarly, some might take on more risk in investing their retirement savings than they would in the absence of a safety net. In such an environment, mandating universal participation simply forces people to pre-pay in the form of social security taxes for benefits they ultimately will receive from the system. Therefore, the purpose of a mandatory system is to protect society against free riders.

While these three arguments explain why governments might believe it important to mandate a minimum level of universal participation in a national retirement programme, they are silent about what the particular level of government benefits should be. These arguments are also silent on whether the government might merely mandate a plan, leaving it to the private sector to manage it. For example, in several countries the other two legs of the retirement-income stool are encouraged by government regulation as an alternative to government provision. Governments often use tax policy to provide incentives for employers and unions to sponsor pension plans that, like the government-run plan, are mandatory and non-assignable. In some of those countries, tax incentives are also given to self-employed individuals and households (who are not otherwise covered) to create a retirement fund for themselves. Use of such funds for other purposes is discouraged by imposing penalties on early withdrawal of money from the fund.

The Role of Occupational Pensions

Pension plans sponsored by employers or unions – also known as occupational pensions – are often integrated with the government-run plan, either explicitly or implicitly. When combined with the government-provided retirement benefit, these plans are usually designed to replace 70–100 per cent of pre-retirement earnings of lower- and middle-income employees in developed nations. Benefits are usually lower for higher-income workers, who then must rely on direct personal savings for a larger part of their retirement income.

Why are employers and/or trade unions logical sponsors of retirement plans for their employees? There are at least four good reasons (Bodie 1990). First, they make for *efficient labour contracting*. Pension plans are an incentive device in labour contracts because they affect employee hiring and turnover patterns, work effort, and the timing of retirement.

Second, they promote *informational efficiencies*. Employment-based plan sponsors often have better access than the plan's beneficiaries to information needed for preparing long-run financial plans tailored to the needs of the employees. In particular, sponsors may have better knowledge of the probable path of future labour income for their employees. By providing a basic plan that saves enough to provide for replacement of anticipated future labour earnings, the corporate sponsor can potentially save more efficiently than each employee acting individually. In order for the sponsor to provide efficiently for future wage and salary replacement of employees, it is enough to have accurate forecasts of the earnings of the group as a whole and not the individual earnings of each member of the group. It is probably easier (although by no means simple) to forecast group earnings than it is to forecast an individual's future earnings.

Third, employment-based plans can avoid *principal–agent problems*. While plan sponsors and beneficiaries may have conflicting economic interests, in many respects their interests coincide. Employers who acquire a reputation for taking care of their employees' retirement needs may find it easier to recruit and retain higher-quality employees. If employees' trust and goodwill towards their employers develop, then motivation and labour productivity may also be enhanced. Employers therefore have some economic incentive to act in the best interests of their employees.

Other possible providers of retirement planning services may be less suitable as beneficial agents of employees. Insurance agents, stockbrokers, and others who are often engaged in providing these services to individual households may be less trustworthy than employers because they could be interested in selling individuals some product or service that those individuals might not choose were they well-informed. These other agents may be motivated to persuade individuals to save too much for retirement or to invest in inappropriate ways. Anyone who has ever tried to find competent and impartial personal financial planning or investment advice is aware of the difficulties.

Fourth, plan sponsors often have *access to capital markets* that is unavailable to their employees acting as individual savers. Employees may not be able to buy certain kinds of insurance individually, but might be able to do so as members of an employee group. In addition, sponsoring firms can take advantage of scale economies while individual employees cannot. Financial intermediaries such as insurance companies can provide a suitable vehicle for the insurance needs of employees. But often a financial intermediary will not be willing to provide enough of the insurance desired by the individual at an efficient price because of problems of adverse selection and moral hazard.

Longevity insurance is an important example of this. In principle longevity risk is diversifiable and can be largely eliminated through risk pooling and sharing. But, as explained earlier, the problem of adverse selection can make the private insurance market for life annuities inefficient. Group insurance through pension plans is often seen as a solution to this problem.

Defined Benefit and Defined Contribution Pension Plans

Pension plans are usually classified in terms of what is promised to the beneficiaries. There are two basic categories: *defined contribution* and *defined benefit* plans. In a defined contribution plan, a formula specifies the amount of money that must be contributed to the plan, but does not specify benefit payouts. Contribution rules are usually a predetermined fraction of salary (for example, the employer contributes ten per cent of the employee's annual wages to the plan), although that fraction need not be constant over an employee's career. The pension fund consists of a set of individual investment accounts, one for each covered employee. Pension benefits are not specified, other than that at retirement the employee gains access to the total accumulated value of the contributions and the earnings on those contributions. These funds can be used to purchase an annuity or can be taken in the form of a lump sum.

In a defined contribution plan, the participating employee frequently has some choice over both the level of contributions and the way the account is invested. In principle, contributions could be invested in any security, although in practice most plans limit investment choices to bond, stock, and money-market funds. The employee retirement account is, by definition, fully funded by the contributions, and the employer has no legal obligation beyond making its periodic contributions. Therefore, in a defined contribution plan much of the task of setting and achieving retirement income replacement goals falls on the employee. In some defined contribution plans, employees have the option of transferring some of the risks to an insurance company.

In a defined benefit plan, by contrast, the pension plan specifies formulae for the cash benefits to be paid after retirement. The benefit formula typically takes into account years of service for the employer and level of wages or salary (for example, the employer pays a retired worker an annuity from retirement to death, the amount of which might be equal to one per cent of his final annual earnings multiplied by years of service). Contribution amounts are not specified, and the employer (called the 'plan sponsor') or an insurance company hired by the sponsor guarantees the benefits and thus absorbs the investment risk. The obligation of the plan sponsor to pay the promised benefits is similar to a long-term debt liability of the employer.

In the United States, the United Kingdom, and many other countries the trend since the mid-1990s has been away from defined benefit towards the defined contribution form. The two plan types are not, however, mutually exclusive. Many sponsors have defined benefit plans as a 'primary' plan, in which participation is mandatory, and supplement them with voluntary defined contribution plans. Moreover, some plan designs are 'hybrids' combining features of both plan types. For example, in a 'cash-balance' plan each employee has an individual account that accumulates interest. Each year, employees are told how much they have accumulated in their account and, if they leave the firm, they can take that amount with them. If they stay until retirement age, however, they receive an annuity determined by the plan's benefit formula. A variation on this design is a 'floor' plan, which is a defined contribution plan with a guaranteed minimum retirement annuity determined by a defined benefit formula. These plan designs usually take into account the benefits provided by the government-run system.

Why Does Funding Matter?

The *pension plan* is the contractual arrangement setting out the rights and obligations of all parties; the *pension fund* is a pool of assets set aside to provide collateral for the promised benefits. In defined contribution plans, the value of the benefits equals that of the assets and so the plan is always exactly fully funded. In contrast, defined benefit plans have a continuum of possibilities. There may be no assets dedicated to the pension plan in a separate fund, in which case the plan is said to be *unfunded*. When there is a separate fund but assets are worth less than the present value of the promised benefits, the plan is *underfunded*. If the plan's assets have a market value that exceeds the present value of the plan's liabilities, it is said to be *overfunded*.

Why and how does funding matter? The assets in a pension fund provide collateral for the benefits promised to the pension-plan beneficiaries. A useful analogy is that of an equipment trust. In an equipment trust, such as one set up by an airline to finance the purchase of airplanes, the planes serve as specific collateral for the associated debt obligation. The borrowing firm's legal liability, however, is not limited to the value of the collateral. By the same token, if the value of the assets serving as collateral exceeds the amount required to settle the debt obligation, any excess reverts to the borrowing firm's shareholders. So, for instance, if the market value of the equipment were to double, this would greatly increase the security of the promised payments, but it would not increase their size. The residual increase in value would accrue to the shareholders of the borrowing firm.

The relation among the shareholders of the firm sponsoring a pension plan, the pension fund, and the plan beneficiaries is similar to the relation among the shareholders of the borrowing firm in an equipment trust, the equipment serving as collateral, and the equipment-trust lenders. In both cases, the assets serving as collateral are 'encumbered' (that is, the firm is not free to use them for any other purpose as long as that liability remains outstanding), and the liability of the firm is not limited to the specific collateral. Any residual or 'excess' of assets over promised payments belongs to the shareholders of the sponsoring firm. Thus the greater the funding, the more secure the promised benefits. However, whether the plan is underfunded, fully funded, or overfunded, the size of the *promised* benefits does not change.

Why do employers fund their defined benefit plans? Reasons appear to vary across countries. First, funding offers benefit security if there is no government insurance of pension benefits, or only partial insurance. Employees may demand that the future pension promises made to them by their employer be collateralized through a pension fund. In the United Kingdom, for example, there is no government pension insurance beyond the minimum guaranteed pension of the State Earnings Related Pension Scheme (SERPS). Pension funding in this case provides an important cushion of safety for retirement income.

Second, some countries impose minimum funding standards by law. These standards seek to insure that promised pension benefits are paid even in the event of default by the corporate sponsor and also aim to protect the government (and the taxpayer) from abuse of government-supplied pension insurance. In the United States, for example, the Pension Benefit Guaranty Corporation (PBGC) must continue pension payments offered by defined benefit pension plans if their sponsoring corporations become bankrupt with an underfunded pension plan. Recent changes in United States pension law mandate that the PBGC insurance premium must depend on the plan's extent of underfunding, and have also eliminated the possibility of voluntary termination of an underfunded pension plan.

Third, there may be tax incentives for plan sponsors to fund their defined benefit plans. Black (1980) and Tepper (1981) have shown that the tax advantage to pension funding stems from the ability of the sponsor to earn the pre-tax rate of return on pension investments. It is no accident that in Germany, where employers face a tax disadvantage if they fund their pension plans, pensions are predominantly unfunded.

Finally, funding a pension plan may provide the sponsoring firm with financial 'slack' that can be used in case of possible financial difficulties the firm may face in the future. In the United States, pension law allows plan sponsors facing financial distress to draw upon excess pension assets by reduced funding or, in the extreme case, voluntary plan termination. The pension fund therefore effectively serves as a tax-sheltered contingency fund for the firm.

Funding of Pensions in the Public Sector

In a strictly unfunded *pay-as-you-go* government-operated pension system, retirees' benefits depend entirely on the stream of revenue generated by taxes levied on currently active workers. If this were exactly true, benefits would fluctuate with changes in economic fortunes, rising when tax collections rose, and falling in recessions. In practice this does not happen because most government pensions are of the defined-benefit variety and promise to deliver retirement benefits according to a specified benefit formula. Nevertheless, without funding, benefit payouts are susceptible to cuts when the public sector experiences a rising ratio of retired to active workers and/or large government deficits. In this event benefits accrued under that formula may be altered as a way of reducing this form of government debt.

As a case in point, consider the 1983 reform of the United States Social Security system. A changing demographic structure for workers led many to become concerned that the future benefits in a pure pay-as-you-go system could be dramatically reduced. Hence, a key provision of that reform was to require substantial pre-funding of future benefits. To do this, the Social Security payroll tax rate was raised and the excess of current revenues over current benefit payments was invested in government bonds held in a trust fund.

While this reform apparently funds the plan, some are less sure about the result. In a private plan, funding is used to insure against default by the plan sponsor. Under Social Security, the promise to pay benefits seemingly has the same level of full faith and credit of the government as the bonds used to fund the plan. Yet there seems to be a belief that pre-funding will ensure that when workers reach retirement they will indeed receive benefits approximating those promised under the current benefit formula (that is, the one in effect when they were active in the labour force).

A problem with this view is that there remains a potential risk associated with benefits promised under a government-run retirement income system. Even if the current government is committed to maintaining the current schedule of promised benefits, it cannot credibly fully bind future governments to do so. Indeed, it has become evident in many countries that the benefit formula and the method of financing those benefits can be and often are changed. In the United States, for example, the Congress has changed both in the past and it can surely do so again in the future. Perhaps more strikingly, public pensions in Chile were radically restructured in the early 1980s, replacing the defined benefit public social security system with a mainly private defined-contribution plan. In the 1990s Australia followed Chile's lead, and several eastern European countries have done so too.

These examples bring out an important difference between government and private-sector obligations. A private-sector plan sponsor cannot unilaterally repudiate its legal liability to make promised payments. It can default because of inability to pay, but it cannot repudiate its legal obligations without penalty. On the other hand, a government – because it has the power to legislate changes in the law – can sometimes find ways to repudiate such obligations without immediate and obvious penalty. Indeed, an integrated system in which private plan sponsors supplement government-provided pension benefits to achieve a promised 'replacement ratio' of pre-retirement earnings can be seen as a type of private-sector insurance against the political risks of the government-run system.

In sum, a mixed public–private system of retirement income provision is a way of reducing the risks of each separate component through diversification across providers. Public-sector pension plans can change the law to reduce promised benefit levels. Private-sector pension plan sponsors are committed by law (and perhaps reputation) to pay promised benefits, but they may default. And sometimes, as an additional linkage reinforcing the first two legs of the retirement income stool, the government may insure private pension benefits against the risk of default (Bodie and Merton, 1993).

See Also

- ▸ Population Ageing
- ▸ Retirement

Bibliography

Black, F. 1980. The tax consequences of long-run pension policy. *Financial Analysts Journal* 36: 25–31.
Bodie, Z. 1990. Pensions as retirement income insurance. *Journal of Economic Literature* 28: 28–49.
Bodie, Z., and E. Davis, eds. 2000. *The foundations of pension finance*. Cheltenham: Edward Elgar.
Bodie, Z., and R. Merton. 1993. Pension benefit guarantees in the United States: A functional analysis. In *The future of pensions in the United States*, ed. R. Schmitt. Philadelphia: University of Pennsylvania Press.
Merton, R. 1983. On consumption indexed public pension plans. In *Financial aspects of U.S. pension system*, ed. Z. Bodie and J. Shoven. Chicago: University of Chicago Press.
Tepper, I. 1981. Taxation and corporate pension policy. *Journal of Finance* 36: 1–13.

Perfect Competition

M. Ali Khan

Abstract

This article attempts a critical appraisal of the literature on perfect competition as it has evolved since the work of Debreu–Scarf and Aumann in the 1960s, following papers of Debreu–Scarf and Aumann. It focuses on mathematical techniques that have been garnered to cope with the presuppositions of the classical theory relating to *finitude, convexity* and *agent-independence.*

Keywords

Additivity; Aggregation; Arrow–Debreu model of general equilibrium; Central limit theorems; Competitive equilibrium; Consumptions sets; Convexity; Core equivalence; Cores; Correspondences; Equilibrium; Existence of general equilibrium; Externalities; Fixed-point theorems; Game theory; Hayek, F. A.; Hyperplanes; Idealized limit economy; Incentive compatibility; Independence; Infinite commodities; Large games; Law of large numbers; Lindahl equilibrium; Lyapunov theorem; Marginal cost pricing; Measure spaces; Negligibility; Non-convexity; Perfect competition; Probability; Public goods; Returns to scale; Shapley–Folkman theorem; Steinitz's theorem; Wald, A.

JEL Classifications

D4

An allocation of resources generated under perfect competition is an allocation of resources generated by the pursuit of individual self-interest and one which is insensitive to the actions of any single agent. Self-interest is formalized as the maximization of profits over production sets by producers and the maximization of preferences over budget sets by consumers, both sets of actions being taken at a price system which cannot be manipulated by any single agent, producer or consumer. An essential ingredient then in the concept of perfect competition, that which gives the adjective *perfect* its thrust, is the idea of *economic negligibility* and, in a set of traders with *many* equally powerful economic agents, the related notion of *numerical negligibility.* Perfect competition is thus an idealized construct akin (say) to the mechanical idealization of a frictionless system or to the geometric idealization of a straight line.

Following the lead of Wald, a mathematical formalization of perfect competition in a setting with an exogenously given finite set of commodities and of agents was developed in the early 1950s in the pioneering papers of Arrow, Debreu and McKenzie. It was shown that convexity and independence assumptions on tastes and technologies guarantee that a competitive equilibrium exists, and that a Pareto- optimal allocation can be sustained as a competitive equilibrium under appropriate redistribution of resources. It was also shown, drawing on the tacit assumption that markets are universal but by avoiding any convexity assumptions, that, with local non-satiation, every competitive allocation is Pareto-optimal. Relegating precise definitions to the sequel, we refer the reader to Koopmans (1961) for a succinct statement of the theory; Debreu (1959) and McKenzie (2002) remain its standard references, Fenchel (1951) and Rockafellar (1970) its mathematical subtexts, and Weintraub (1985), Ingrao and Israel (1987) and Mirowski (2002) its sources of historical appraisal.

However, in its exclusive focus on drawing out the implications of convexity and agent-independence for a formalization of perfect competition, the theory remained silent about environments with increasing marginal rates, in production and in consumption, as well as those where private and social costs and benefits do not coincide, to phrase this silence in Pigou's (1932) vocabulary of a preceding period. In particular, the notion of perfect competition that was fashioned by the initial theoretical development had no room for economic phenomena emphasized, for example, in the papers of Hotelling (1938), Hicks (1939) and Samuelson (1954). It took around two decades to show that, at least as far as collective consumption and public goods were concerned, the theory had within it all the resources for an elegant incorporation, but of course within the confines and limitations of its purview (see Foley 1970 and his followers). Non-convexities in production and consumption were a different matter entirely; they required mathematical tools that went beyond convexity, and further development had to await the invention of non-smooth calculus of Clarke and his followers; see Rockafellar and Wets (1998) and Mordukhovich (2006) for a comprehensive treatment.

A robust formalization of the idea of perfect competition for non-convex technological environments in the specific form of marginal cost pricing equilibria, with the regulation of the increasing returns to scale producer(s) given an explicit emphasis, can be outlined under each of the three headings of the theory identified by Koopmans: existence and the two welfare theorems. Marginal cost pricing equilibria exist under suitable survival and loss assumptions, but are not globally Pareto-optimal even under the assumption of universality of markets. Finally, Pareto-optimal allocations can be sustained as marginal cost-pricing equilibria under appropriate redistribution of resources. Moreover, under the terminology of Lindahl–Hotelling equilibria, Khan and Vohra (1987) provide the existence of an equilibrium concept that incorporates both public goods and increasing returns to scale in one sweep. This work on perfect competition in the presence of individualized prices stemming from collective consumption and a regulated production sector (or sectors) merits an entry in its own right, and rather than a detailed listing of the references, we refer the reader to Vohra (1992) and Mordukhovich (2006, ch. 8) for details and references.

Three observations in connection with this recent, but already substantial, literature are worth making. First, in the attempts to generalize the second fundamental theorem of welfare, one can discern a linguistic turn whereby both the Arrow–Debreu emphasis on decentralization and the Hicks–Lange–Bergson–Samuelson–Allais equality of marginal rates are seen as special cases within a synthetic treatment emphasizing the intersection of the cones formalizing marginal rates; Khan's (1988) introduction is an emphatic articulation of this point of view. Second, a canonical formulation of the notion of marginal rates, despite fits and starts, now seems within reach, though a notion that works well for the necessary conditions may not be the one equally suited for the question of existence; see Hamano (1989) and Khan (1999). Finally, conceptual clarity requires an understanding of circumstances when this type of non-convex theory bears a strong imprint of its finite-dimensional, convex counterpart, as detailed in Khan (1993), as opposed to when its higher reaches require a functional-analytic direction totally different from that charted out in the pioneering papers of the 1950s; see Bonnisseau and Cornet (2006) for reference to recent work.

With price-taking assumed rather than endogenously deduced, there is no overriding reason why a formalization of perfect competition must limit itself to a setting with a finite, as opposed to an unbounded (infinite), number of (perfectly divisible) commodities. Indeed, another set of pioneering papers of Debreu, Hurwicz and Malinvaud, written in the 1950s with an eye to a theory of intertemporal allocation but over a time horizon that is not itself arbitrarily given, fixed and finite so to speak, did consider the decentralization of efficient production plans as profit - maximizing ones. But again, it was only two decades later that the work of Bewley, Peleg-Yaari, Gabszewicz and Mertens inaugurated sustained attempts to provide a general formalization of perfect competition over infinite-dimensional commodity spaces (see Khan and Yannelis 1991). The work can again be categorized under Koopmans's three headings of the theory, but relative to its finite-dimensional counterpart, it noted that the separation of disjoint convex sets, and the use of aggregate resources to furnish a bound on the consumption sets to ensure compactness, proved to be matters of somewhat greater subtlety. In short, even a norm-compact set of an infinite-dimensional commodity space is 'rather large' and its cone of non-negative elements 'rather small'. Indeed, as Negishi's method of proof attained dominance, the imbrication of the convexity assumption in a clear demarcation of fixed-point theorems for issues of existence and separating hyperplane theorems for those of decentralization, no longer obtained. The subject is surveyed in Mas-Colell and Zame (1991), but another survey is perhaps overdue as exploration of individual mathematical structures, ordered structures in particular, reveals hitherto unforeseen essentials, and increasing returns to scale and other nonclassical phenomena are inevitably accommodated; see the references of Aliprantis et al. (2002, 2006), on the one hand, and those of Shannon (1999) and Bonnisseau (2002) on the other.

However, the question persists as to what meaning can be given to the study of perfect competition in a setting with an exogenously given infinite-dimensional commodity space where markets open only once and there is no room for the correction of mistakes and unfulfilled plans. If the extension of the theory requires additional technical assumptions, how do they translate into *desiderata* that are of relevance for the formalization of the coherence of decentralized, self-interested decision-making of independent agents acting independently of each other? Even if, for example, the *uniform properness* assumption of Mas-Colell (1986) and his followers could be pinned down as a formalization of bounded marginal rates of substitution (see the notion of a Fatou cone in Araujo et al. 2004, and one failed attempt in Khan and Peck 1989), what does it say about the set-up of the model itself that lifts this up to be a limitation as fundamental as that of convexity or independence? If the underlying motivation for the extension to infinite-dimensional commodity spaces is time, risk, quality, information or location, how do these considerations manifest themselves in the infinite dimensionality of the commodity space, in a situation that necessities (or precludes) one commodity in an economy being numerically negligible relative to the entire set? More sharply, why ought not the resulting problems be more squarely faced in simpler partial equilibrium models, rather than studied under the limitation of a construction whose primary emphasis is the viability and desirability of static interaction? We defer these issues to turn to our principal theme, namely, the formalization of the perfectness of perfect competition.

The point is that the assumption of a finite number of agents embodied in all of this work is an explicit admission of the fact that the economic non-negligibility of each agent, at least in principle, and therefore her non-manipulation of, and corresponding submission to, the price system furnishes a somewhat muted maximization of her self-interest. In terms of the emphasis on *negligibility* as a prerequisite for a rigorous formalization of perfect competition, as is being emphasized in this article, the postulated behaviour of individual agents in the so-called Arrow–Debreu–McKenzie model of perfect competition, with or without infinite commodities, externalities and increasing returns to scale, leads to the rather natural puzzlement as to what it is precisely that guarantees an agent's passive acceptance of the price system, let alone individualized pricing rules, and that too in a construction whose primary motivation is consistency and generality. In the vernacular due to Hurwicz (1972), one that has gained increasing currency since the 1980s, what is it that makes this model of the economic system *incentive- compatible*? How is its gloss of the intuitive notions of *negligibility, large* and *many* to be made precise?

Six conceptually separate attempts to answer this question are distinguished here; these alternative but interrelated formalizations of perfect competition draw their meaning from two early conjectures: (i) Edgeworth's (1881) conjecture on the shrinking of the *core* to its set of competitive allocations (again, precise definitions to follow), and (ii) Farrell's (1959) conjecture on the existence of competitive equilibrium in a environment that is not necessarily convex. Interpreted literally, both conjectures are clearly false for a given finite economy, but the first can be distinguished from the second in not being simply a case of dispensing with an assumption in a result whose basic contours are well-established, but rather in going beyond Koopmans's categorization of perfect competition to include a solution concept other than that of Pareto optimality. It is in the reliance of the core notion as a test for the perfectness of competition, in working with a third fundamental theorem of welfare economics, so to speak, and in giving precision to the ambiguity inherent in the term *shrinking*, that allows an entry into the formalization of the negligibility of individual agents. However, at this point, the discussion demands the rigour of notation and definitions; and since the essence of the ideas can be adequately communicated in the context of an economy without producers, that is, in an *exchange economy*, we confine ourselves to this case.

An *exchange economy* consists of a commodity space L, a set of traders T, a space of trader characteristics P defined on the commodity space, and a mapping $\mathscr{E}$ from T into P with the value of $\mathscr{E}$ at a particular t in T being given by the triple $\mathscr{E}(t) = ((X(t), \succcurlyeq_t, e(t))$ specifying the characteristics of agent t in T. The space of characteristics is thus a product space constituted by consumption sets $X(t) \subseteq L$, by binary relations $\succcurlyeq_t$ over $X(t) \times X(t)$, preferences over the consumption set read 'preferred or indifferent to', and by initial endowments $e(t) \in X(t)$. An *allocation* x: $T \rightarrow L$ is an assignment of commodity bundles such that $x(t) \in X(t)$ for all t in T and such that the summation, suitably formalized, of $(x(t) - e(t))$ over T is zero, or, in the case of *free-disposal*, less than or equal to zero. In either case, the fundamental economic problem facing a particular exchange economy, as discussed above and being given symbolic formulation here, is the choice of an allocation.

An allocation x: $T \rightarrow L$ is said to be in the *core* if there does not exist any other allocation y and a coalition S $\subseteq$ T, suitably formalized, such that $y(t) \succcurlyeq_t x(t)$ and not $x(t) \succcurlyeq_t y(t)$ for all $t \in S$, and that the summation of $(y(t) - e(t))$ over S is zero, or again with free disposal, less than or equal to zero. A perfectly competitive allocation of resources is a price-based allocation where a *price system* is a non-zero, continuous linear function on the commodity space L. A *competitive equilibrium* is a pair (p, x) where p is a price system and x an allocation such that for all t in T, $x(t)$ is a maximal element for $\succcurlyeq_t$ in the budget set $\{y \in X(t)) : (y, p) \preccurlyeq (e(t), p)\}$. Here (y, p) denotes the valuation of the commodity bundle y by the function p and, in case L is the Euclidean space IR^{ℓ}, the Riesz representation theorem allows it to be given a simple accounting interpretation of an inner product $(y, p) = \Sigma_{i=1}^{\ell} p_i y_i$; see Rudin (1974) for this theorem and for other unspecified terminology. For any competitive equilibrium (p, x), x is referred to as a *competitive allocation.* In terms of the earlier discussion of infinite-dimensional commodity spaces, the commodity space L has presumed on it enough mathematical structure so as to give meaning to the ordering 'less than or equal to', to the summation operator in the notion of an allocation and of a blocking coalition, and to linearity and continuity in the notion of a price system. Conceptually, what is of consequence here is that competitive allocations can be viewed as making precise the idea of some sort of individual rationality, and core allocations as making precise the idea of some sort of group rationality.

In Aumann's (1964) formulation of perfect competition, the set of traders is the Lebesgue unit interval, the commodity space is the Euclidean non-negative orthant IR^{ℓ}_{+}, the set of admissible coalitions the Borel σ-algebra on the unit interval, and summation, Lebesgue integration. Under the assumption of Lebesgue measurability of preferences $\succcurlyeq_t$, and of Lebesgue integrability of the initial endowments $e(\cdot)$, he proved that the set of competitive allocations of such an economy coincides with its set of core allocations and, in Aumann (1966), that neither set is empty. These precise and elegant affirmations of the conjectures of Edgeworth and Farrell did not require any convexity hypotheses on preferences, and, what is perhaps of equal significance, they furnished a precise formulation of an idealized limit economy in which price-taking is rendered theoretically reputable: every agent is numerically and economically negligible in that the effect of his or her action, not only on the price system but also on the equilibrium allocation, is precisely zero. An agent has a negligible weight very much akin (say) to the probability of a particular point on a dartboard being hit by a dart.

The seminal nature of Aumann's conception was quickly realized and incorporated in to the mainstream. The metaphor of a continuum of agents is now routinely (but not incorrectly) invoked to validate the removal of idiosyncratic uncertainty by aggregation even in models of a representative agent in theoretical work in macroeconomics and other, so-called applied, fields. This work that is nothing if not an investigation of competition, perfect or otherwise. Two observations are worth making. First, whereas Aumann's assumption of a Lebesgue unit interval was only a simplifying one, and that the results hold for any arbitrary *atomless* and finite measure space, Lebesgue (rather than Riemann–Stieltjes) integration is essential to a theory based on T as the set of agent-names, and therefore free of any topological considerations; see Khan and Sun (2002, Introduction) for a detailed exposition of this point. Indeed, Shapley has even questioned the postulate of measurability, leave alone continuity, for a notion of an allocation whose very *raison d'être* is a formalization of independent individual self-interest. Second, since the theory is based on a neglect of sets of measure zero, it is a conception of an allocation as an equivalence class of functions, rather than of functions themselves, that is identified by the theory. Put more sharply, Pareto-optimal allocations in an economy with a continuum of agents do not exist if their definition is taken verbatim from that of a finite economy, and not recast in terms of coalition of positive measure. In any case, the theory of an economy E conceived as a measurable map, at least in its finite-dimensional embodiment, is a testimony to the power of the Lyapunov theorem on the range of an atomless vector measure and to a powerful mathematical theory of the integration of correspondences that emerges as its corollary; Hildenbrand (1974) is the relevant reference.

A contemporaneous formulation of Vind (1964) short-circuits some of these issues concerning sets of zero measure by ignoring agents altogether, and focusing instead on coalitions, each with its own preferences and endowments, as the primitive data of the economy. Allocations then are measures on a non-atomic measure-space, and the notions of core and competitive allocations, correspondingly defined, can be shown to be identical solution concepts. This is a formulation of perfect competition that is also measure-theoretic, but one, alternative to that of Aumann, that explicitly does away with mathematical integration as its necessary microfoundation. However, by assuming countable additivity, Vind enabled Debreu (1967) to draw on Radon–Nikodym differentiation to effect a reconciliation. It took subsequent work of Armstrong and Richter to give fuller autonomy to this alternative point of view by first eliminating countable additivity, and then in setting the discussion in the framework of nonatomic Boolean algebras; see Armstrong and Richter (1986) and their references. Whereas the technical underpinning of this approach is now clearly seen to be the Armstrong and Prikry (1981) extension of the Lyapunov theorem, it is perhaps fair to say that the conceptual ramifications of this alternative (perhaps syndicalist) vision have yet to be fully explored and understood; see Avallone and Basile (1998) and Basile and Graziano (2001) for references to current research.

The formulation of perfect competition due to Brown and Robinson (1975), the third to be discussed here, returns to the methodological individualism of Aumann, and requires the set of agent-names T to be an internal star-finite set, the commodity space to be ${}^*\mathrm{IR}^{\ell}_{+}$, the nonstandard extension of IR^{ℓ}_{+} based on manipulable infinitely large and infinitesimally small numbers, the summation in the definitions of allocations and core to be summation over internal sets, the set of admissible coalitions to be the set of all internal subsets of T and $\mathscr{E}$ to be an internal map from T to ${}^*\mathscr{P}$, the set of agent characteristics modelled on ${}^*IR^{\ell}_{+}$. Such a formulation utilizes methods of nonstandard analysis, a specialization in mathematical logic due to A. Robinson; see Loeb and Wolff (2000) for details and references. On replacing equality by equality modulo infinitesimals in the definitions of allocation and the core, Brown and Robinson (1975) and, without their ad hoc standardly bounded assumption on allocations, Brown and Khan (1980) showed the equivalence (and Brown 1976, and Khan 1975, the existence) of core and competitive allocations of a nonstandard economy without any convexity assumptions on preferences. Loeb's (1973) combinatorial analogue of Lyapunov's theorem provided the mathematical underpinning of the theory. This alternative affirmation of the conjectures of Edgeworth and Farrell is another way of making precise the concepts of *many* agents and of their individual *negligibility:* meaning can be given to an individual trader's actions having a positive, but infinitesimal, effect on the price system and on an allocation. Even though an initial motivation of this work was to explore a formulation of perfect competition and of a *large* economy in a vernacular alternative to that of measure theory, it was heavily influenced by measure-theoretic formulations, but with an added emphasis on asymptotic implementation (discussed below), something clear even in the earliest papers of Brown–Robinson and Khan; see Rashid (1987), Anderson (1991) for details and references.

Relative to the classical theory brought to a culmination by Arrow, Debreu, McKenzie, Uzawa, Gale, Nikaido and Negishi, and succinctly surveyed in Koopmans (1961), the literature discussed above can be read as an exploration of the structural analytics of the set of agents of a stylized economy. Where Aumann takes the replicated sequence of Debreu–Scarf to a countably additive atomless measure space of agents, Brown–Robinson take it to a star-finite internal set each of whose points (agents) is given the same weight, and Armstrong–Richter, following Vind's cue, to a finitely-additive atomless measure space of coalitions. A fourth direction, intriguing and not yet fully synthesized in and with the other three, is represented in the work of Kaneko and Wooders (1986, 1989) and Hammond et al. (1989); also see Hammond (1995), Kaneko and Wooders (1994, 1996), Winter and Wooders (1994) and their references. The heart of this approach is to grapple with absolute and proportional magnitudes within the same framework, to focus on finite coalitions chosen from a continuum, through the notion of a *measure-consistent* partition. It concerns an atomless countably additive measure space of agents in which a single agent (and therefore a finite set of agents) is closed and thereby measurable, and a set of measure-preserving isomorphisms. A notion of an *f*-core is formulated and shown to be equivalent to the set of competitive equilibria *even* with externalities, and to the so-called Aumann core, without externalities; Wooders (1997) focuses on public goods. This approach yields its own particular way of looking at the continuum as a idealized limit of a finite economy, one that revolves around finer and finer measure-consistent partitions of an atomless continuum. It is thereby different in spirit from the more conventional way that asymptotic implementation has been formalized. We refer the reader to the references for details, and turn to what is seen here as the fifth formulation of perfect competition.

Strange as it may seem in retrospect, the idealizations of Aumann and Brown–Robinson were criticized on grounds of realism, on the observation that there do not exist economies with uncountably many agents; see Koopmans (1974) and the Georgescu-Roegen–Rashid exchange discussed in Khan (1998). The work categorized here as an asymptotic implementation of the idealized limiting versions of perfect competition was motivated, in part, by this criticism (ironically also used by Armstrong–Richter as their stated motivation for finitely additive measures), and, in part, by a methodological curiosity as to whether the results established for nonstandard and measure-theoretic economies are artifacts of the way *negligibility* and *large* economies were being modelled. Taking its point of departure from the replicated sequences of Debreu and Scarf (1963), the response is to consider a sequence $\mathscr{G} = \{\mathscr{E}_k\}_{k=1}^{\infty}$ of finite economies based on the commodity space IR_+^l where $\mathscr{E}_k$ is an economy with a set of agents T_k of cardinality k. For each finite economy $\mathscr{E}_k$, competitive and core allocations can be defined in the conventional way without encountering any technical difficulties in the formalization of summation or of a coalition. It is clear that agents in $\mathscr{E}_k$ get increasingly numerically negligible with an increase in k, and given a uniformly bounded assumption on initial endowments, also get increasingly economically negligible. For this *perfectly competitive sequence* of economies, one can ask: for any $\varepsilon > 0$, however small, does there exist an integer k_o such that core allocations of all $\mathscr{E}_k \in \mathscr{G}, k \succcurlyeq k_0$, can be sustained as approximate competitive equilibria, and whether such equilibria exist, with ε indicating in either instance, the degree of approximation? In short, are the formulations of perfect competition in idealized limit economies capable of an asymptotic implementation, with an arbitrarily fine degree of approximation, in economies of arbitrarily large but finite cardinality?

Asymptotic equivalence and existence theorems under varying degrees of generality followed quickly once the problem was posed. We shall not touch upon the various elaborations and refinements except to note that they have been obtained under two disparate techniques, both drawing on the results for an idealized limit economy. The first, associated especially with Hildenbrand, is to conceive of an economy as a measure on the space of characteristics and to utilize Skorokhod's theorem and the theory of weak convergence of measures on a topological space (typically metrizable) of characteristics $\mathscr{P}$. Under Debreu's rather vivid terminology of 'neighboring economic agents', such topologies were formulated by Debreu, Kannai, Hildenbrand–Mertens, Grodal and others, and surely have independent interest; see Hildenbrand (1974). The second approach is based on the observation that 'any sentence which is true in the standard universe is true for internal entities in the nonstandard universe', and as such, results pertaining to a nonstandard exchange economy can be 'flipped over', as it were, to a corresponding result for a large but finite economy. The differences between the two approaches are interesting from a methodological point of view: the fact that one approach is, in principle, not inherently dependent on any topology on the space of preference relations or on their continuity (as in Khan and Rashid 1976, 1982) and applies as readily to core as to competitive allocations (as in Khan 1974), suggests a further look as to how the other may be extended; see Anderson (1992) for a comprehensive treatment. In any case, we have two mutually supporting ways of extracting information for large but finite economies from idealized limit economies, even of the mixed type with atoms that generated the scepticism about idealized limit economies in the first place; see Gabszewicz and Shitovitz (1992) and their references. This claim is further underscored by a development due to Loeb (1975), but before turning to it, we discuss what may be seen as fifth formulation of *negligibility* and thereby of perfect competition.

The asymptotic interpretation of the perfectness of perfect competition concerns sequences of economies, and a question arises as to whether, given an arbitrary economy rather than an arbitrary degree of approximation, one can find the error, independent of the number of agents, with which the equivalence and existence theorems hold. Thus, rather than ask how large is large enough, one asks how small is small enough for the assumption of price-taking behaviour to be unjustified. For the question posed in this way, initially by Starr (1969), it was the definitive result of Anderson (1978) that capped initial explorations of Arrow–Hahn, Henry, Shaked and others. With the shedding of compactness and continuity assumptions under the nonstandard approach, Anderson observed that the argument in Khan and Rashid (1976) could be based on the Shapley–Folkman theorem instead of that of Loeb (1973) (itself based on Steinitz's theorem), and carried out entirely in standard terms to obtain an elementary equivalence theorem. This yields the asymptotic results as corollaries, and also furnishes them with a rate of convergence, a consideration emphasized by Shapley (1975). The same observation applied to Khan and Rashid (1982) led to an elementary existence theorem; see Geller's (1986) extension of Anderson et al. (1982).

In the prominence that it gives to a fixed finite economy, this sixth and final fifth formulation of perfect competition connects directly to the results whose introduction began this entry; it emphasizes that the equalities in the results surveyed by Koopmans, and the counter-examples implicitly underlying them, perhaps ought to be given a probabilistic cast rather than taken completely literally. In his alternative proof of the Shapley–Folkman theorem, Cassels (1975) had already emphasized this connection. Mas-Colell deepened it further by appealing to results of especial sophistication concerning the law of large numbers and the central limit theorem, and by noting that his refinement of the equivalence theorem has 'no analogue in Aumann's continuum of traders model', and that the precise probabilistic estimates that this approach offers have no counterpart in the continuum framework (see Anderson 1992, Sections 8 and 9 for details and precise references). However, it is undeniable that it is the exact results for the idealized limit economies that generally indicate the directions of pursuit of the approximations for a finite economy: approximations and numerical algorithms come into play once the exact has been exactly identified. Thus, from a substantive point of view, modulo fine technicalities, how a particular issue pertaining to perfect competition is set, measure-theoretic or nonstandard or asymptotic, is largely a contextual matter of analytical convenience and preference.

This conclusion is further sharpened by the methodological unification offered in Loeb (1975) (see Khan and Sun 1997b, for exposition). It is the central claim of this article that Loeb probability spaces go a long way towards settling the question of how the perfectness of perfect competition is to be given a precise mathematical formulation. It is already clear in Aumann's pioneering papers that perfect competition draws from the atomlessness rather than any other particularities of the measure space of agents: the metric on the unit interval, or the topology of any topological measure space, is not, indeed cannot be, of any direct relevance. What is presumably of the essence is that the space of agents' names be hospitable to measurability as well as to independence (the latter term now being used in its precise probabilistic sense rather than as a reference to an absence of externalities), that it generates results capable of straightforward asymptotic implementation, and that, for concepts that revolve only on distributions of the allocations as in Hart–Kohlberg, it yields solutions that are insensitive to a permutation of agent names. In the context of *large* games (discussed below), Khan and Sun (1996, 1999b) make the case for Loeb spaces on the basis of these *desiderata* and emphasize their dual identity in the 'pushing down' and 'lifting up' theorems: being standard, measure spaces, any result on an abstract measure-space (Aumann) economy applies to them, and thereby to an internal non-standard (Brown–Robinson) economy and hence can be asymptotically interpreted; or alternatively, any approximate result can be translated, as indicated above, to a non-standard economy, and thereby pushed down to its standard Loeb measure-theoretic counterpart. As such, Loeb spaces go a considerable way in obliterating the sixfold categorization of perfect competition that marks this entry.

Going beyond method to mathematical substance, atomless Loeb spaces are ideally suited for operations ensuring that aggregation removes the irregularities that arise from non-convexities as well as from idiosyncratic uncertainty. In a systematic and far-reaching development, Sun established that the integrals and distributions of correspondences defined on Loeb spaces and taking values in a separable infinitedimensional Banach space, in the first instance, and into Polish spaces (separable and completely metrizable) in the second, have all the properties that the theory of perfect competition requires of them. Moreover, a perfectly satisfactory law of large numbers for a continuum of random variables is obtained, and for a such a continuum, the notions of *independence* and of *exchangeability* are dual in a very elegant sense, and yields, as in Duffie and Sun (2007), the existence of an independent random matching. Supplementing the notion of an economy as a random variable, the measurability of the map noted above, a stochastic economy can now be formalized as a stochastic process on a product space, the space of agent names T and an atomless Loeb space of states of nature, Ω, to reveal circumstances under which the distributions of core and competitive allocations of a sampled economy coincide, or approximately coincide in the case of a large economy, with those of the deterministic (population) economy; see Sun (1999). Further application of this substantial theory is noted below; here the reader is referred to Sun's chapters in Loeb and Wolff (2000, chs. 7 and 8) for exposition and full mathematical references. (For references to work on random economies that does not rely on Loeb spaces, see Radner 1982, Section 7.6, and Majumdar and Rotar 2000.)

In taking stock at this stage, we underscore the fact that even though six robust and logically related methods of studying perfect competition have been illustrated through the conjectures of Edgeworth and Farrell, the discussion could, in principle, equally well have been conducted through alternative tests based on alternative solution concepts: the value (Hart 2002 and his references), or the bargaining set (Anderson 1998 and his references), or Cournot's conjecture (Mas-Colell 1986; Novshek and Sonnenschein 1983 and their references), all now conceived in a setting where individual agents are negligible. Alternatively, we could discuss applications, particularly in mathematical finance where Arrow markets and ideas of negligibility find concrete expression in derivative financial instruments and in *well-diversified* portfolios (see Anderson and Raimondo 2006; Khan and Sun 1997a, respectively, for references). However, rather than turn to them and make this article unmanageable, we draw on the rich and diverse formulation of perfect competition at our disposal to consider the substantive issues broached earlier: public goods, externalities, increasing returns to scale and infinite commodities, all under the rubric of static interaction. Ironically, non-convexities in idealized limit economies have concerned consumptions sets and survival assumptions rather than increasing returns to scale technologies (see Trockel 1984; Hammond 1993 and their references); research efforts have been most active in the study of public goods and externalities, and here the theory dovetails, from a technical point of view, into work on infinitedimensional commodity spaces.

The formalization and defence of perfect competition has, from the very beginning, proceeded on the independence assumption: the fact that individual agents are not related other than through the price system, with a 1952 paper of McKenzie's being the sole exception. Thus Hayek (1948, pp. 96–7) quotes Stigler in emphasizing the 'explicit and complete exclusion from the theory . . . of all personal relationships existing between the parties'. Such relationships are *external* to the perfected concept, and, to the extent that positive and normative content can be cleanly distinguished, externalities, and the Pigovian private–social divergences that they entail, have strong and negative implications for its normative content. If the nonconvexities identified by Starrett (1972) are ignored (but also see Otani and Sicilian 1977), Arrow's universality requirement for the first fundamental theorem of welfare economics can always be met by the creation of markets, fictitious or otherwise, but it clearly leans on a particularly acute form of myopia. Arrow securities and Lindahl prices for public goods, and more generally, prices for contingent commodities, and personalized prices for more pervasive externalities, bring out an obvious tension between incentive compatibility and efficiency. As emphasised in Starrett (1971), if there is a commodity that reflects a particular agent's dependence on my consumption, why should she or I, let alone the others, take the price of that commodity to be given and non-manipulable? or take myself to be economically negligible?

Of course, one response to these difficulties is to face the future as a future without the fiction of a complete set of commodity markets or, equally imaginatively, existing markets for securities that span *all* contingencies. Under this alternative, one can regard the price system itself as a means of fostering a relationship between the parties, and to conceive of a rationality which explicitly incorporates the informational resources of the *others* in the economy. This is to look on a price system as an instrument of solidarity as well as an instrument of allocation, a keeping up with the Joneses, not so much in their actions as in the individualized information that undergirds their actions, a move reminiscent of Veblen in the space of information rather than that of conspicuous consumption. This is a move inaugurated by Radner (1967), and it leads to a notion of equilibrium, a *rational expectations equilibrium* so to speak, in which both aspects of the price system are taken into account while not necessarily departing from the purview of the static Arrow–Debreu–McKenzie theory. One can only wonder what mathematical form such a theory will take when it is set in the framework in which individual agents are *negligible;* we point the reader to Radner (1982) and Jordan and Radner (1982) for details and references, and revert to the idealized limit economy.

There is also a technical problem in the consideration of pervasive externalities in an idealized limit economy. Since the individualistic, as opposed to the coalitionally based, approach to perfect competition works with an equivalence class of functions from the space of agent-names to agent-actions rather than the function itself, it is difficult to give meaning to one agent's dependence on the actions of another. In a context of a Lindahl equilibrium of an idealized limit economy, even one with a finite number of commodities and a single public good, one has to reckon with the fact that public goods enjoin equality instead of aggregation, and thereby force the analysis out of a finite-dimensional Euclidean space, as in the Aumann–Brown–Robinson limit theory, to a search for a suitably tractable space of equivalence classes of functions of individualized prices. It is these attendant functional-analytic difficulties, perhaps as much as the fact that the incentive- compatibility problems are most acute in this setting, that have discouraged the initial exploratory attempts of Roberts, Emmons and Khan and Vohra from being followed up; see Khan and Vohra (1985) for references. And it is precisely difficulties of this kind that also prevent a successful theory for idealized limit economies with non-ordered preferences; see Balder's (2000) interpretive use of the argument in Khan and Papageorgiou (1987), originally due to Grodal, to turn a positive proof into a negative claim of inconsistency, a claim that apparently derails the initial exploration of Khan and Vohra (1984) and their followers. Externalities, rather than being widespread, need to be controlled and confined in an idealized limit economy. This previous sentence, as well as the tone of this entire paragraph so far, runs counter to the fourth approach to perfect competition associated with Hammond, Kaneko and Wooders, but, as emphasized above, the integration of this fourth approach with the other five has not yet been fully achieved. The theory is under active development, and it is too early to say that a formulation sufficiently robust as to be deemed canonical has been achieved (see Balder 2007b; Cornet and Topuzu 2005; Hammond 1995; Kaneko and Wooders 1994; Noguchi 2005; Noguchi and Zame 2006 and their references).

In its dissociation of the study of perfect competition from its roots in welfare theory, the inclusion of externalities makes explicit its connection to game theory. Competitive equilibria with externalities take their place next to marginal-cost pricing and Cournot–Nash equilibria in violating Pareto optimality, but do allow one to ask whether decentralized self-interested decision-making is consistent in the aggregate if it is taken with respect to certain measurable indices of societal responses rather than solely with respect to a price system. Such a formulation of perfect competition goes back to the early 1950s in the papers of McKenzie and Debreu, and to the 1970s in Chipman's formulations of Marshallian parametric externalities. Indeed, the original proof of Arrow–Debreu of the existence of competitive equilibrium revolved around viewing the economy as a game in which the only 'personal relationship' between the parties relates to that with a fictitious auctioneer, a point of view that finds fuller expression in the Shafer–Sonnenschein notion of an *abstract economy.* In more recent investigations of a *large game*, the literature takes another turn towards probability theory, and conceives of an agent's actions as resulting from maximization that takes as given the distribution, or individual moments, of the random variable summarizing societal responses. The question then reduces to the existence of such equilibrium distributions, but with social interaction, however limited, recourse has to be made to assumptions on ideal types, and on the *conditional* or *mutual independence* of these types (see the prescient remarks of Hayek 1948, p. 47). This is a theory of competition in which Loeb spaces, and the Dvoretsky–Wald–-Wolfowitz extension of the Lyapunov theorem play a dominant role; see Khan and Sun (1999b, 2002); Khan et al. (2006) Loeb and Sun (2006) and their references to the work of Schmeidler, Radner–Rosenthal, Milgrom–Weber and Mas–Colell. (Balder (2007a) offers a perspective based on Young measures.)

The technical machinery forged through the study of large games enables a broadened notion of economic negligibility, one that includes *informational negligibility* in an environment with asymmetric information. In a 1936 article on 'Economics and Knowledge', Hayek (1948, pp. 43–44) had already supplemented Adam Smith's emphasis on the *division of labour* by the principle of the *division of knowledge* and asked.

> whether, in order that we can speak of equilibrium, every single individual must be right, or whether it would not be sufficient if, in consequence of a compensation of errors in different directions, quantities of the different commodities coming on the market were the same as if every individual had been right. A fuller discussion of this problem would have to consider the whole question of the significance which some economists (including Pareto) attach to the law of great numbers in this connection.

The issue is: 'right' about what? The problem devolves on anticipations and expectations, beliefs about beliefs regarding each other and the price system, and it does not require more than a mild degree of scepticism to abandon fictional markets responding to predetermined and universally agreed upon states of nature. There is a need for viable notions of independence and aggregation to eliminate idiosyncratic risk and nullify 'combination of fragments of knowledge existing in different minds'. Sun (2006) and Sun and Yannelis (2007a, 2007b) give pride of place to the Fubini property in idealized limit economies, and consolidate earlier applications of Loeb spaces for a successful resolution of Malinvaud's work on insurance markets, and that of Gul, McLean and Postlewaite on the compatibility of efficiency and incentive compatibility; also see Jackson and Manelli (1997). Khan and Sun (1999a) and Sun (2006) also present compelling arguments why finitely additive measures and the conventional product measure cannot respond to the technical difficulties.

The problems arising from asymmetric information are, at their root, problems of agent interdependence that cannot be internalized through markets, and as such represent particularly recalcitrant externalities; the assumptions that Sun–Yannelis impose on their signal process can be seen as one successful attempt to subdue them. And in an idealized limit economy with *many* commodities, each commodity seen on its own rather than through the externalities' lens, one has to cope with the fact that Lyapunov's theorem is false for an infinite dimensional vector measure, in addition to all of the problems discussed earlier. It is the *thinness* of its target space, as proposed by Kingman–Robertson in the late 1960s, that allows an atomless probability space of agents to work its magic in the form of the existence and equivalence theorems; see Kluvanek and Knowles (1976) and Diestel and

Uhl (1977) for necessary and sufficient conditions for the validity of the Lyapunov theorem. There is a hidden assumption, to adopt the postmodern flourish of Tourky and Yannelis (2001), in the Aumann–Brown–Robinson formulations of perfect competition, and the equivalence theorem can fail when the qualitative relationship between the cardinalities of agents and commodities fails; in addition to Muench's example, see Forges et al. (2001) and Serrano et al. (2001). More generally, if the intricacies of reaching binding agreements in coalition formation cannot be bracketed away, how can a concept embodying group rationality coincide with one hinging on individual rationality? An option, but one that goes against the very grain of this article, is to dissociate competition from price-taking entirely and derive it as a consequence, as in the no-surplus characterizations of Makowski and Ostroy (2001, Section 9) and Serrano and Volij (2000). The field is under active development; in addition to the papers of Sun, Tourky and Yannelis, see Forges et al. (2002), Herves-Beloso et al. (2005), Martins-da-Rocha (2003, 2004), and Podczeck (1997, 2001, 2004) and their references.

In his classic 1936 tour de force, Hayek deconstructed the Arrow–Debreu–McKenzie construction before it was constructed, so to speak, by distinguishing between an a priori 'pure logic of choice' and an empirical science. In so far as this article, in its focus on existence and core equivalence, has concentrated on the adjective *perfect,* and avoided questions of cardinality, computability, learning and stability of a perfectly competitive allocation of resources, it has neglected the noun *competition* as being outside its scope. For this, the reader could perhaps begin with Morgan (1993), and move from there to Arrow (1986), Buchanan (1987) and Radner (1991), and from there, if she is still so inclined, to the entire gamut of economic theory.

See Also

- ▶ Competition
- ▶ Large Economies
- ▶ Measure Theory
- ▶ Non-Standard Analysis
- ▶ Shapley–Folkman Theorem

Bibliography

Aliprantis, C.D., B. Cornet, and R. Tourky. 2002. Economic equilibrium: Optimality and price decentralization. *Positivity* 6: 205–241.

Aliprantis, C.D., M. Florenzano, and R. Tourky. 2006. Production equilibria. *Journal of Mathematical Economics* 42: 406–421.

Anderson, R.M. 1978. An elementary core equivalence theorem. *Econometrica* 46: 1483–1487.

Anderson, R.M. 1991. Non-standard analysis with applications to economics. In *Handbook of mathematical economics*, ed. W. Hildenbrand and H. Sonnenschein, vol. 4. New York: North-Holland.

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

Anderson, R.M. 1998. Convergence of the Aumann–Davis–Maschler and Geanakopolos bargaining sets. *Economic Theory* 11: 1–37.

Anderson, R.M., M.A. Khan, and S. Rashid. 1982. Approximate equilibria with bounds independent of preferences. *Review of Economic Studies* 49: 473–475.

Anderson, R.M., and R. Raimondo. 2006. *Equilibrium in continuous-time financial markets: Endogenously dynamically complete markets*, Working paper. Berkeley: University of Berkeley.

Araujo, A., V.F. Martins-da-Rocha, and P. Monteiro. 2004. Equilibria in reflexive Banach lattices with a continuum of agents. *Economic Theory* 24: 469–492.

Armstrong, T.E., and K. Prikry. 1981. Liapunoff's theorem for nonatomic, finitely additive, bounded, finite-dimensional vector-valued measures. *Transactions of the American Mathematical Society* 266: 499–514. Erratum in the same journal (1982), 272, 809.

Armstrong, T., and M.K. Richter. 1986. Existence of nonatomic core-Walras allocation. *Journal of Economic Theory* 38: 137–159.

Arrow, K.J. 1986. Economic theory and the hypothesis of rationality. *Journal of Business* 59: S385–S399.

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

Aumann, R.J. 1966. Existence of competitive equilibria in markets with a continuum of traders. *Econometrica* 34: 1–17.

Avallone, A., and A. Basile. 1998. Lyapunov–Richter theorem in B-convex spaces. *Journal of Mathematical Economics* 30: 109–118.

Balder, E.J. 2000. Incompatibility of usual conditions for equilibrium existence in continuum economies without ordered preferences. *Journal of Economic Theory* 93: 110–117.

Balder, E.J. 2007a. Comments on purification in continuum games. *International Journal of Game Theory* 37.

Balder, E.J. 2007b. More on equilibria in competitive markets with externalities and a continuum of agents. *Journal of Mathematical Economics* 44.

Basile, A., and M.G. Graziano. 2001. Restricted coalition formation mechanisms in finitely-additive economies. *Journal of Mathematical Economics* 36: 219–240.

Bonnisseau, J.M. 2002. The marginal pricing rule in economies with infinitely many commodities. *Positivity* 6: 275–296.

Bonnisseau, J.M., and Cornet, B. 2006. *Existence of equilibria with a tight marginal pricing rule*. Cahier de la MSE, Université Paris 1.

Brown, D.J. 1976. Existence of competitive equilibrium in a non standard exchange economy. *Econometrica* 44: 537–547.

Brown, D.J., and M.A. Khan. 1980. An extension of the Brown–Robinson equivalence theorem. *Applied Mathematics and Computation* 6: 167–175.

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

Buchanan, J.M. 1987. *Economics: Between predictive science and moral philosophy.* College Station: Texas A&M University Press.

Cassels, J.W.S. 1975. Measures of the non-convexity of sets and the Shapley–Folkman–Starr theorem. *Mathematical Proceedings of the Cambridge Philosophical Society* 78: 433–436.

Cornet, B., and M. Topuzu. 2005. Existence of equilibria for economies with externalities and a measure space of consumers. *Economic Theory* 26: 397–421.

Debreu, G. 1959. *The theory of value*. New York: Wiley.

Debreu, G. 1967. Preference functions on measure spaces of economic agents. *Econometrica* 35: 111–122.

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

Diestel, J., and J.J. Uhl. 1977. *Vector measures*. Providence: American Mathematical Society.

Duffie, D., and Y.N. Sun. 2007. Existence of independent random matching. *Annals of Applied Probability* 17: 386–419.

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

Farrell, M.J. 1959. The convexity assumption in the theory of competitive markets. *Journal of Political Economy* 67: 377–391.

Fenchel, H. 1951. *Convex cones, sets and functions*, Lecture notes. Princeton: Princeton University Press.

Foley, D. 1970. Lindahl's solution and the core of an economy with public goods. *Econometrica* 38: 66–72.

Forges, F., A. Heifetz, and E. Minelli. 2001. Incentive compatible core and competitive equilibria in differential information economies. *Economic Theory* 18: 349–365.

Forges, F., E. Minelli, and R. Vohra. 2002. Incentives and the core of an exchange economy: A survey. *Journal of Mathematical Economics* 38: 1–41.

Gabszewicz, J., and B. Shitovitz. 1992. The core in imperfectly competitive economies. In *Handbook of game theory with economic applications*, ed. R.J. Aumann and S. Hart, vol. 1. Amsterdam: North-Holland.

Geller, W. 1986. A improved bound for approximate equilibria. *Review of Economic Studies* 53: 307–308.

Hamano, T. 1989. On the non-existence of the marginal cost pricing equilibrium and the Ioffe normal cone. *Zeitschrift für National Ökonomie* 50: 47–53.

Hammond, P.J. 1993. Irreducibility, resource-relatedness, and survival in equilibrium with individual nonconvexities. In *General equilibrium, growth and trade II*, ed. R. Becker, M. Boldrin, R. Jones, and W. Thomson. New York: Academic Press.

Hammond, P.J. 1995. Four characterizations of constrained Pareto efficiency in continuum economies with widespread externalities. *Japanese Economic Review* 46: 103–124.

Hammond, P.J., M. Kaneko, and M.H. Wooders. 1989. Continuum economies with finite coalitions: Core equilibria and widespread externalities. *Journal of Economic Theory* 49: 113–134.

Hart, S. 2002. Values of perfectly competitive economies. In *Handbook of game theory with economic applications*, ed. R.J. Aumann and S. Hart, vol. 3. Amsterdam: North-Holland.

Hayek, F.A. 1948. *Individualism and economic order.* Chicago: Chicago University Press.

Herves-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.

Hicks, J.R. 1939. The foundations of welfare economics. *Economic Journal* 49: 696–712. Also, Prefatory note to the 1984 reprint in *Wealth and welfare: Collected essays in economic theory*, vol. 1. Oxford: Basil Blackwell.

Hildenbrand, W. 1974. *Core and equilibria of a large economy.* Princeton: Princeton University Press.

Hotelling, H. 1938. The general welfare in relation to problems of taxation and of railway and utility rates. *Econometrica* 6: 242–269.

Hurwicz, L. 1972. On informationally decentralized systems. In *Decision and organization*, ed. C.B. McGuire and R. Radner. Amsterdam: North-Holland.

Ingrao, B., and G. Israel. 1987. *La Mano invisibile.* Roma-Bari: Gius. Laterza & Figli Spa. English translation, *The invisible hand: Economic equilibrium in the history of science.* Cambridge: MIT Press, 1990.

Jackson, M., and A. Manelli. 1997. Approximate competitive equilibria in large economies. *Journal ofEconomic Theory* 77: 354–376.

Jordan, J., and R. Radner. 1982. Rational expectations in microeconomic models: An overview. *Journal of Economic Theory* 26: 201–223.

Kaneko, M., and M.H. Wooders. 1986. The core of a game with a continuum of players and finite coalitions: The model and some results. *Mathematical Social Sciences* 12: 105–137.

Kaneko, M., and M.H. Wooders. 1989. The core of a continuum economy with widespread externalities and finite coalitions: From finite to continuum economies. *Journal of Economic Theory* 49: 135–168.

Kaneko, M., and M.H. Wooders. 1994. Widespread externalities and perfectly competitive markets: Examples.

In *Imperfections and behavior in economic organizations*, ed. R.P. Gilles and P.H.M. Ruys. Boston: Kluwer.

Kaneko, M., and M.H. Wooders. 1996. The nonemptiness of the *f*-core of a game without side-payments. *International Journal of Game Theory* 25: 245–258.

Khan, M.A. 1974. Some remarks on the core of a 'large' economy. *Econometrica* 42: 633–642.

Khan, M.A. 1975. Some approximate equilibria. *Journal of Mathematical Economics* 2: 63–86.

Khan, M.A. 1988. Ioffe's normal cone and the foundations of welfare economics: An example. *Economics Letters* 28: 15–19.

Khan, M.A. 1993. Lionel McKenzie and the existence of competitive equilibrium. In *General equilibrium, growth and trade II*, ed. R. Becker, M. Boldrin, R. Jones, and W. Thomson. New York: Academic Press.

Khan, M.A. 1998. *Representation, language and theory: Georgescu-Roegen on methods of economic science*. Paper presented at the Colloque International L'Oeuvre Scientifique de Nicholas Georgescu-Roegen, Strasbourg, 7–9 November.

Khan, M.A. 1999. The Mordukhovich normal cone and the foundations of welfare economics. *Journal of Public Economic Theory* 1: 309–338.

Khan, M.A., and N. Papageorgiou. 1987. On Cournot–Nash equilibrium in generalized qualitative games with an atomless measure space of players. *Proceedings of the American Mathematical Society* 100: 505–510.

Khan, M.A., and N.T. Peck. 1989. On the interiors of production sets in infinite dimensional spaces. *Journal of Mathematical Economics* 18: 29–39.

Khan, M.A., and S. Rashid. 1976. *Limit theorems on cores with costs of coalition formation*, Johns Hopkins working paper No. 24. Abridged version published as 'A limit theorem for an approximate core of a large but finite economy', *Economics Letters* 1 (1978): 297–302.

Khan, M.A., and S. Rashid. 1982. Approximate equilibria in markets with indivisible commodities. *Journal of Economic Theory* 28: 82–101.

Khan, M.A., and Y.N. Sun. 1996. Non-atomic games on Loeb spaces. *Proceedings of the National Academy of Sciences of the United States of America* 93: 15518–15521.

Khan, M.A., and Y.N. Sun. 1997a. The capital-asset-pricing model and arbitrage pricing theory: A unification. *Proceedings of the National Academy of Sciences of the United States of America* 94: 4229–4232.

Khan, M.A., and Y.N. Sun. 1997b. On Loeb measure spaces and their significance for non-cooperative game theory. In *Current and future directions in applied mathematics*, ed. M. Alber, B. Hu, and J. Rosenthal. Berlin: Birkhäuser.

Khan, M.A., and Y.N. Sun. 1999a. Weak measurability and characterizations of risk. *Economic Theory* 13: 441–460.

Khan, M.A., and Y.N. Sun. 1999b. Non-cooperative games on hyperfinite Loeb spaces. *Journal of Mathematical Economics* 31: 455–492.

Khan, M.A., and Y.N. Sun. 2002. Non-cooperative games with many players. In *Handbook of game theory with economic applications*, ed. R.J. Aumann and S. Hart, vol. 3. Amsterdam: North-Holland.

Khan, M.A., and R. Vohra. 1984. Equilibrium in abstract economies with out ordered preferences and with a measure space of agents. *Journal of Mathematical Economics* 13: 133–142.

Khan, M.A., and R. Vohra. 1985. On the existence of Lindahl equilibria in economies with a measure space of non-transitive consumers. *Journal of Economic Theory* 36: 319–332.

Khan, M.A., and R. Vohra. 1987. Lindahl–Hotelling equilibria. *Journal of Public Economics* 34: 143–158.

Khan, M.A., and N.C. Yannelis. 1991. *Equilibrium theory in infinite dimensional spaces*. New York: Springer.

Khan, M.A., K.P. Rath, and Y.N. Sun. 2006. The Dvoretzky–Wald–Wolfowitz theorem and purification in atomless finite-action games. *International Journal of Game Theory* 34: 91–104.

Kluvanek, I., and G. Knowles. 1976. *Vector measures and control systems*. Amsterdam: North-Holland.

Koopmans, T.C. 1961. Convexity assumptions, allocative efficiency, and competitive equilibrium. *Journal of Political Economy* 69: 478–479.

Koopmans, T.C. 1974. Is the theory of competitive equilibrium with it? *American Economic Review* 69: 325–329.

Loeb, P.A. 1973. A combinatorial analog of Lyapunov's theorem for infinitesimally generated atomic vector measures. *Proceedings of the American Mathematical Society* 39: 585–586.

Loeb, P.A. 1975. Conversion from nonstandard to standard measure spaces and applications in probability theory. *Transactions of the American Mathematical Society* 211: 113–122.

Loeb, P.A., and Y.N. Sun. 2006. Purification of measure-valued maps. *Illinois Journal of Mathematics* 50: 747–762.

Loeb, P.A., and M. Wolff. 2000. *Nonstandard analysis for the working mathematician*. Dordrecht: Kluwer.

Majumdar, M., and V. Rotar. 2000. Equilibrium prices in a random exchange economy with dependent agents. *Economic Theory* 15: 531–550.

Makowski, L., and J.M. Ostroy. 2001. Perfect competition and the creativity of the market. *Journal of Economic Literature* 39: 479–535.

Martins-da-Rocha, V.F. 2003. Equilibria in large economies with a separable Banach commodity space and non-ordered preferences. *Journal of Mathematical Economics* 39: 863–889.

Martins-da-Rocha, V.F. 2004. Equilibria in large economies with differentiated commodities and non-ordered preferences. *Economic Theory* 23: 529–552.

Mas-Colell, A. 1986. The price equilibrium existence problem in topological vector lattices. *Econometrica* 54: 1039–1054.

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. 4. Amsterdam: North-Holland.

McKenzie, L.W. 2002. *Classical general equilibrium theory*. Cambridge: MIT Press.

Mirowski, P. 2002. *Machine dreams*. Cambridge: Cambridge University Press.

Mordukhovich, B.S. 2006. *Variational analysis and generalized differentiation, I & II*. Berlin: Springer.

Morgan, M. 1993. Competing notions of 'competition' in late nineteenth-century American economics. *History of Political Economy* 25: 563–604.

Noguchi, M. 2005. Interdependent preferences with a continuum of agents. *Journal of Mathematical Economics* 41: 665–686.

Noguchi, M., and W.R. Zame. 2006. Competitive markets with externalities. *Theoretical Economics* 1: 143–166.

Novshek, W., and H. Sonnenschein. 1983. Walrasian equilibria as limits of noncooperative equilibria. *Journal of Economic Theory* 30: 171–187.

Otani, Y., and J. Sicilian. 1977. Externalities and problems of nonconvexity and overhead costs in welfare economics. *Journal of Economic Theory* 14: 239–252.

Pigou, A.C. 1932. *The economics of welfare*, 4th ed. London: Macmillan. 1st edn, 1920.

Podczeck, K. 1997. Markets with infinite commodities and a continuum of agents with non-convex preferences. *Economic Theory* 9: 385–426.

Podczeck, K. 2001. Core and Walrasian equilibria when agents' characteristics are extremely dispersed. *Economic Theory* 22: 699–725.

Podczeck, K. 2004. On Core-Walras equivalence in Banach spaces when feasibility is defined by the Pettis integral. *Journal of Mathematical Economics* 40: 429–463.

Radner, R. 1967. Equilibre des marchés á terme et au comptant en cas d'incertitude. *Cahiers d'econométrie* 9: 30–47.

Radner, R. 1982. Equilibrium under uncertainty. In *Handbook of mathematical economics*, ed. K.J. Arrow and M.D. Intriligator, vol. 2. New York: North-Holland.

Radner, R. 1991. Intertemporal general equilibrium. In *Value and capital: Fifty years later*, ed. L.W. McKenzie and S. Zemagni. London: Macmillan.

Rashid, S. 1987. *Economies with many agents*. Baltimore: Johns Hopkins Press.

Rockafellar, R.T. 1970. *Convex Analysis*. Princeton: Princeton University Press.

Rockafellar, R.T., and J.-B. Wets. 1998. *Variational analysis*. New York: Springer.

Rudin, W. 1974. *Real and complex analysis*. New York: McGraw-Hill Book Co..

Samuelson, P.A. 1954. The pure theory of public expenditures. *Review of Economics and Statistics* 36: 387–389.

Serrano, R., and O. Volij. 2000. Walrasian allocations without price-taking behavior. *Journal of Economic Theory* 95: 79–106.

Serrano, R., R. Vohra, and O. Volij. 2001. On the failure of core convergence with asymmetric information. *Econometrica* 69: 1685–1696.

Shannon, C. 1999. Increasing returns in infinite horizon economies. *Review of Economics Studies* 64: 73–96.

Shapley, L.S. 1975. An example of a slow-converging core. *International Economic Review* 16: 345–351.

Starr, R.M. 1969. Quasi-equilibria in markets with nonconvex preferences. *Econometrica* 37: 25–38.

Starrett, D.A. 1971. A note on externalities and the core. *Econometrica* 41: 179–183.

Starrett, D.A. 1972. Fundamental nonconvexities in the theory of externalities. *Journal of Economic Theory* 4: 180–199.

Sun, Y.N. 1999. The complete removal of individual uncertainty: Multiple optimal choices and random economies. *Economic Theory* 14: 507–544.

Sun, Y.N. 2006. The exact law of large numbers via Fubini extension and the characterization of insurable risks. *Journal of Economic Theory* 126: 31–69.

Sun, Y.N., and N.C. Yannelis. 2007a. Perfect competition in asymmetric information economies: Compatibility of efficiency and incentives. *Journal of Economic Theory* 134: 175–194.

Sun, Y.N., and N.C. Yannelis. 2007b. Core, equilibria and incentives in large asymmetric information economies. *Games and Economic Behavior* 61 (1): 131–155.

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.

Trockel, W. 1984. *Market demand: An analysis of large economies with nonconvex preferences*. Berlin: Springer.

Vind, K. 1964. Edgeworth-allocations in an exchange economy with many traders. *International Economic Review* 5: 165–177.

Vohra, R. 1992. Marginal cost pricing under bounded marginal returns. *Econometrica* 60: 859–876.

Weintraub, E.R. 1985. *General equilibrium analysis: Studies in appraisal*. Cambridge: Cambridge University Press.

Winter, E., and M.H. Wooders. 1994. An axiomatization of the core for finite and continuum games. *Social Choice and Welfare* 11: 165–175.

Wooders, M.H. 1997. Equivalence of Lindahl equilibrium with participation prices and the core: An axiomatization of the core for finite and continuum games. *Economic Theory* 9: 115–127.

Perfect Foresight

Margaret Bray

Keywords

Cobweb model; Expectations; Imperfect information; Intertemporal general equilibrium; Overlapping generations models; Perfect foresight; Probability distribution; Rational expectations equilibrium; Stationary state; Uncertainty

JEL Classifications
D4

Perfect foresight is an occasionally convenient theoretical assumption whose total lack of realism is undisputed, and perhaps unrivalled. There are two elements to perfect foresight; firstly that people have definite point expectations, allowing no uncertainty, of future variables, and secondly that these expectations are correct. In practice, as these fortunate perfectly foresightful individuals generally inhabit models with instantaneously clearing perfectly competitive markets, they only need to forecast prices. The pioneering work by Hicks (1939) on intertemporal general equilibrium theory provides a framework in which the issues associated with perfect foresight can be explored. Writing prior to the development of the expected utility theory of choice under uncertainty (von Neumann and Morgenstern 1944), Hicks had no alternative to a deterministic model in his discussion. He acknowledges the existence and importance of uncertainty in expectation formation, but argues in a somewhat unsatisfactory fashion that point predictions can be interpreted as risk-adjusted summaries of underlying probability distributions. Hicks divides time into weeks. Trade takes place weekly. Supply and demand in each week depend upon decisions made in the past, expectations of spot prices in future weeks, and current spot prices. In temporary equilibrium these spot prices adjust to clear markets, but expectations may be wrong. In the situation which Hicks terms ‘Equilibrium over Time’, markets clear at each date, and, crucially, everyone has perfect foresight; price expectations are fulfilled.

Hicks’s insight that perfect foresight is an equilibrium concept is important. If people have non-equilibrium expectations, the temporary equilibrium prices in the current spot markets differ from the prices in full equilibrium over time, and the effects of current investment and production decisions based on mistaken expectations reverberate through the future. This can be illustrated in the simplest model of supply and demand in which expectations play a part: the cobweb model, used by Kaldor (1934) in discussing disequilibrium adjustments, and by Muth (1961) in the paper which gave us the phrase ‘rational expectations’. In the cobweb model, demand at t D_t, depends upon the price at t p_t, $D_t = a - bp_t$. Supply depends upon point expectations p_t^e formed before t about $p_t; S_t = cp_t^e$. In temporary equilibrium supply equals demand, $a - bp_t = cp_t^e$, so $p_t = \left(a - cp_t^e\right)/b$. In the perfect foresight equilibrium expectations are correct $p_t^e = p_t = a/(b + c)$. If the price is and has been at the perfect foresight equilibrium level for a long time people will, quite reasonably, expect this price to persist. In an economy in a long-run stationary state with unchanging prices perfect foresight is plausible. Difficulties arise when a shift in an exogeneous variable changes the perfect foresight equilibrium price. Suppose that in the cobweb model an increase in costs causes the supply curve to shift to $S_t = c'p^e$. If people are aware of the change, and understand fully the working of their economy, they may at once calculate and expect the new equilibrium price; alternatively they may all believe the forecast generated by the brilliant economist who knows it all. Less well-informed people may be forced to use past prices in forming their expectations. If these expectations are not at the new equilibrium value $a/(b + c')$ actual prices also differ from equilibrium prices; the economy will take some time to adjust to its new equilibrium and may, as Kaldor shows, fail to get there at all. The dynamic adjustment process, as people try to learn from their mistakes, depends very much upon how they learn, and is not understood in any generality (Bray 1983).

As Hicks argued, equilibrium over time with perfect foresight is most plausible when people expect prices to remain steady, and they do remain steady at the expected level. In the long-run stationary state with no uncertainty there is no need to distinguish between current prices and price expectations. Supply and demand can be thought of as relating to either. In this context the atemporal textbook theory of production and consumption can be reinterpreted to describe a world where production takes time, and is determined by price expectations as well as prices.

In the long-run stationary state tastes and technology must be unchanging and the size of the population and supplies of natural resources static, or possibly in a semi-stationary state growing steadily. These conditions are demanding and implausible. Further, they are not always sufficient for steady prices. As Grandmont (1985) shows, a very simple overlapping generations model has a constant price equilibrium, but may have other perfect foresight equilibria in which the price follows a very complicated, possibly chaotic, path. Unless people know precisely the underlying nonlinear difference equation generating prices they may have great difficulty in inferring prices from past prices.

Postulating perfect foresight allows another reinterpretation of an atemporal general equilibrium model to allow for time (Debreu 1959). In the atemporal model there is a list of different commodities, and a market and price for each commodity. These markets all operate simultaneously; in general equilibrium they all clear. The same mathematical formalism can be used to describe an intertemporal model, by distinguishing commodities by their date of delivery as well as by their characteristics. Commodities may be produced or consumed at a number of different dates, but all trade takes place at the initial date, in a complete set of spot and contingent futures markets. This of course strains credibility; only a very limited number of futures markets exist. However, as Bliss (1975) shows, the same trades, production and consumption can take place if there is a futures market for one good at each date and spot markets for all other goods, provided everyone foresees the full equilibrium over time prices perfectly. There is little to be gained in realism by exchanging the myth of complete markets for the fantasy of perfect foresight. The value of this approach lies in the handle which the well-understood atemporal general equilibrium theory gives in seeking to understand those aspects of intertemporal economics where mistakes in expectation formation appear unimportant.

The most obvious limitation of perfect foresight models is the absence of uncertainty; but the concept has been extended to allow for uncertainty, in the form of the 'rational expectations hypothesis'. This allows expectations to take the form of a probability distribution rather than a point, and requires the distribution to be correct. This begs the question of what is meant by a correct probability distribution. In a theoretical model this is conceptually straightforward. Writing down a theoretical model quite naturally generates a probability distribution describing people's beliefs about certain variables, and another describing the actual probability distribution of these variables. In a rational expectations equilibrium these are the same. In simple cases it may be easy to show that a rational expectations equilibrium exists, by solving the equations equating the distributions.

Consider, for an example, a slight generalization of the cobweb model discussed earlier, in which demand $D_t = a - bp_t + \varepsilon_t$ where ε_t is a normal random variable with mean 0 and variance 1. The price $p_{t.}$ is now a random variable; suppliers believe that it is normal with mean $\widehat{p}_t^e$ and variance $\widehat{\sigma}_t^2$ and want to supply $S = c\widehat{p}_t^e - \widehat{\sigma}_t^2$. In temporary equilibrium supply equals demand, $a - bp_t + \varepsilon_t = c\widehat{p}_t^e - \widehat{\sigma}_t^2$ so $p_t = \left(a - c\widehat{p}_t^e - \widehat{\sigma}_t^2 + \varepsilon_t\right)/b$. Given the $N(0,\ 1)$ distribution of ε_t the price p_t is indeed normally distributed with mean $Ep_t = \left(a - c\widehat{p}_t^e - \widehat{\sigma}_t^2\right)/b$ and variance $1/b^2$. The suppliers have rational expectations if $Ep_t = \widehat{p}_t^e$ and $\widehat{\sigma}_t^2 = 1/b^2$ in which case $Ep_t = \widehat{p}_t^e = \left(a + 1/b^2\right)/(b + c)$.

In more complex theoretical models the mathematics is more difficult, but the concept is clear enough. But is it plausible? The very name 'rational expectations equilibrium' is based on the presumption that this is how rational, optimizing economic agents form expectations. This requires, minimally, that they should, at some point, be able to tell whether their beliefs are correct or not. Apart from examples of the card-choosing or coin-tossing type which have little economic relevance, empirical knowledge of the probability distribution of, for example, a price, depends upon repeated observations of that price. If the probability distribution is stationary, given enough observations of the price, the statistical frequency distribution of past prices reveals the underlying probability distribution. But stationarity is a very strong condition to require. Even if the exogenous random variable (ε_t in the example) is stationary, the distribution of p_t will change as beliefs change. As noted above, we know very little about dynamic adjustment processes outside perfect foresight or rational expectations equilibria.

Knight (1921) uses the term ‘risk’ to describe situations where probabilities can be inferred from data giving the results of repeated observations of similar events, or symmetry arguments (for example coin-tossing). He reserves ‘uncertainty’ for situations concerning unique events where there is no such basis for numerical probability assessments. It is a matter of some philosophical debate whether it is in fact possible to interpret probability numerically in situations which Knight calls uncertainty; subjectivists claim that it is possible, but make no claim that different people will make the same probability assessments. Whatever the outcome of this debate the rational expectations hypothesis is in trouble in situations of Knightian ‘uncertainty’ because there is no single ‘correct’ probability distribution.

Knight argues that economies with risk, but no uncertainty, are essentially identical to economies with perfect foresight, whereas uncertainty (which he claims is all pervasive in business decisions) has a very great effect on the workings of the economy, accounting for imperfect competition and the existence of profit. Risk is unimportant because its effects are nullified by the ability to hedge, to diversify through stock markets, and most importantly because all risks can be perfectly insured. In the light of more recent theory, Knight is clearly wrong, but his argument anticipates recent developments in a fascinating way.

The formalism of the Arrow–Debreu model can be extended to allow for risk and uncertainty, as well as time, by assuming a complete set of contingent futures markets. Commodities are distinguished by the contingencies in which they are available as well as the date. This provides complete insurance. This model has all the properties of the Arrow–Debreu model without risk (existence of equilibrium and Pareto efficiency); thus far Knight’s intuition is correct. Knight is also correct in his observation that in practice complete insurance is not available for many contingencies; we do not live in a world of complete markets. His grand theme is that the presence of uncertainty as opposed to risk renders complete insurance impossible; but in his detailed discussion ‘moral hazard’ plays a key role. Moral hazard is due to the incentive insurance gives to take less care to avoid accidents, and explains why complete insurance is rarely available. As Knight points out, it is a very widespread phenomenon; any implicit or explicit contract which allows one of the parties discretion whose exercise cannot be observed by the other is subject to moral hazard. It is, as Knight argues, all-pervasive in business. But it does not require uncertainty in Knight’s sense; if there is risk and imperfect information there is moral hazard. The economics of information has been an enormously active area of theoretical research in recent years; considerable progress has been made by formal modelling of situations with imperfect information, giving us a much clearer view of its considerable importance and implication. We know that these make for an economics which is qualitatively quite different from that of the Arrow–Debreu model. We would not have learnt this if theorists had not been willing to make assumptions which cannot be taken literally or completely defended, in order to pursue questions. Quantitative probability, perfect foresight and rational expectations have been crucial tools in developing our understanding of economics.

See Also

▶ Uncertainty and General Equilibrium

References

Bliss, C.J. 1975. *Capital theory and the distribution of income*. Amsterdam: North- Holland.

Bray, M.M. 1983. Convergence to rational expectations equilibrium. In *Individual forecasting and aggregate outcomes*, ed. R. Frydman and E.S. Phelps. Cambridge: Cambridge University Press.

Debreu, G. 1959. *Theory of value: An axiomatic analysis of economic equilibrium. Coules foundation monograph no. 17*. New York: Wiley.
Grandmont, J.-M. 1985. On endogenous competitive business cycles. *Econometrica* 53: 995–1045.
Hicks, J.R. 1939. *Value and capital*. Oxford: Clarendon Press.
Kaldor, N. 1934. A classificatory note on the determinateness of equilibrium. *Review of Economic Studies* 1: 122–136.
Knight, F.H. 1921. *Risk, uncertainty and profit*. Boston: Houghton Mifflin.
Muth, J.F. 1961. Rational expectations and the theory of price movements. *Econometrica* 29: 315–335.
von Neumann, J., and O. Morgenstern. 1944. *Theory of games and economic behavior.* Princeton: Princeton University Press.

Perfect Information

Leonard J. Mirman

Keywords

Arrow–Debreu model of general equilibrium; Common knowledge; Competitive equilibrium; Complete information; Cooperation; Cooperative games; Cores; Cournot–Nash equilibrium; General equilibrium; Partial equilibrium; Perfect information; Price system; Stationary state; Tâtonnement; Uncertainty

JEL Classifications

D4

Perfect information is usually thought of as complete knowledge of a person's economic environment. It is clear that nobody in a real economy has perfect knowledge about every aspect of the economy. However it has been argued that perfect knowledge is unnecessary since the price system summarizes all necessary information. Under this line of reasoning the only information that economic agents need are their own tastes and prices. This seems like a very naive argument. However, the real world is more complicated than this argument suggests. Even the prices system itself is not so simple: there are nonlinear prices, for example quantity discounts, as well as different prices for exactly the same commodity. Moreover the economy would function quite differently if the information structure was different, for example if all agents had more knowledge about economic variables. Hence the question arises: how are prices and information used in ideal models of the economy where many very complicated real world relationships have been simplified? In the following discussion the effect of information and the value of prices in conveying and summarizing this information in economic models is described. It appears that in economic models of the economy the 'information content' of prices is not as valuable as it appears on the surface. A well-functioning economy needs much more information than is contained in the price system.

In the quest for the effect of information on the economic environment two basic models come to mind. These are the general equilibrium and partial equilibrium models. The remarks in this article will be aimed basically at the Walrasian general equilibrium model without production. However, many of the points dealing with equilibrating prices – and information – can be made about partial equilibrium models as well as general equilibrium models with production.

The Walrasian paradigm envisioned an economy consisting of a large number of agents trading many goods. Each person, at each point in time, knowing their own tastes and stock of resources (or endowments) decides how much of each good to buy or sell at each possible price, that is, excess demands can be calculated on the basis of each person's environment (tastes and endowments) and the market price. Walras envisioned a steady state or stationary economy. Prices were thought to be generally in equilibrium, known to the consumers or economic agents but with perhaps slight, insignificant fluctuations. In this stable environment the price system regulates the supply to the market and dictates market clearing. In fact each individual agent reacts to the price system which summarizes all necessary information for this agent. Hence if any agent found himself confronted with an equilibrium price vector, then knowledge of only his own tastes and endowments would be sufficient to find the demands which equilibrate the market. However, to actually find the equilibrium price vector requires considerably more information. This difficulty also occurs in a partial equilibrium environment in which the price regulates the market clearing quantity and even the long-run number of producers.

This perception of the Walrasian economy translates into the well-known modern general equilibrium model. In this model there are generally assumed to be a finite number of commodities and a finite number of agents. Each agent is assumed to be a price taker. Equilibrium is found from the market clearing condition on the basis of the aggregate excess demand functions. The price taking behaviour is somewhat unnatural with a finite number of agents since each agent has some market power and therefore would naturally be expected to use strategic behaviour rather than passively taking prices as given. However, there is a very important extension of this model in which there is a continuum of agents. In this extension price taking behaviour is natural since no agent has any market power.

In this economy agents, in deriving their excess demands, need only information about their own tastes and endowments as well as prices. This model is analogous to the Walrasian paradigm described above. However, the analogy does not hold exactly since in the mathematical model there is no historical equilibrium price vector. Hence it is necessary to use an agent outside of the model, for example an auctioneer, to set equilibrium prices. In order to do this, aggregate excess demands must be known to the auctioneer. As a result, although each agent needs to know prices only, any equilibrating mechanism can work only if it has information about all the agents. If no auctioneer is used then it is necessary to design some sort of tâtonnement or groping mechanism to find equilibrating prices. However for such a mechanism to work and to converge to equilibrium prices, it must take account of all agents, In particular, information about excess demands must be available to make the mechanism work.

In order to make clear exactly how information is used in an economic environment, consider a simple economy consisting of two goods and two individuals. Each agent is assumed to take prices as given. For each of these agents only the price is required to describe excess demands while knowledge of both consumers is needed for an equilibrium. Suppose now that agent 2 can have two possible endowments. The first possibility corresponds to a good year while the second corresponds to a lean year. The good year results in high endowments and the lean year results in low endowments. The process of determining the excess demand function for agent 1 remains the same as before; no knowledge of agent 2 is necessary. Agent 2 on the other hand has two possible excess demand functions – one corresponding to the good year, the other to the lean year. These two excess demand functions will in general be quite different, leading to two quite different equilibria. Since he is a price taker no knowledge of agent 1 is needed by agent 2. To find equilibrium prices and allocations, however, the exact characteristics of each agent must be known, no matter what means is chosen to find an equilibrium.

The ideas discussed above can be illuminated by studying various equilibrium concepts and their informational requirements from the theory of games. In particular the information requirements in the general equilibrium model can be highlighted using the core concept of a cooperative game.

First consider the various notions of information in the game context. A distinction is made between games with perfect information and games with complete information. Perfect information in the game theoretic sense pertains to knowledge of the previous history of the game; that is, for perfect information all previous actions of the agents and equilibrium outcomes of the game are known. The notion of complete information in a game theoretic setting pertains to knowledge about the environment. In the general equilibrium context, complete information means that each agent knows his own taste and endowments as well as the tastes and endowments of all other agents. An even sharper notion of information is used in game theoretic models. This is the notion of common knowledge. Common knowledge implies not only that each agent knows his own environment – complete information – but each agent knows that the other agents know that the first agent has complete information and so on ad infinitum.

To see the importance of the common knowledge requirement in a noncooperative game consider a duopolistic market structure using the Cournot–Nash equilibrium concept. In this model each firm maximizes profits given the behaviour of the other firm. An equilibrium is a pair of outputs which is optimal for each firm given that the other firm is playing its equilibrium strategy (or output). In this model, each firm must know its own and its opponents' payoff function but each firm must also know that the opponent knows this information. This is clearly the case since the opponent's strategy will depend upon whom he thinks he is playing against. Moreover the opponent should know that the first firm knows that the opponent has this information. This chain must be continued indefinitely in order to achieve a Cournot–Nash equilibrium. Clearly for a Cournot–Nash equilibrium to obtain, that is, for the common knowledge requirement to be valid, a great deal of information is required.

Another game theoretic equilibrium concept is the core of an economy. The general equilibrium model is a very natural setting for the cooperative notion of the core. The relationship between the purely game-theoretic idea of the core and the general equilibrium concept using prices again illustrates the importance and role of information in a Walrasian general equilibrium model. The core of a general equilibrium economy is defined as the set of outcomes or allocations which cannot be improved upon by any coalition or group of agents. This means that, for any allocation in the core, no subset of agents can band together, trade among themselves using their own endowments and make each agent as well off and at least one agent better off than with the allocation in the core. The core is a cooperative game with complete information. Since the idea of a core involves coalitional or cooperative behaviour the core and competitive equilibrium are quite different. In particular the price taking assumption is incompatible with cooperative behaviour. Hence it is not surprising that more information seems to be needed to find the set of core allocations. The surprising result is that for economies with a continuum of players the set of core allocations coincide with the set of competitive allocations. The use of a continuum of agents is a natural way to model price taking behaviour since no individual agent has power to affect prices. The notion of a core for large economies involves the use, by each agent, of considerably more information than the competitive economy, and yet for large economies the informational content of both notions is exactly the same. Moreover even for finite economies a similar, although not identical, statement can be made. This result is surprising since the core does not contain any explicit reference to prices. However, the relationship between competitive equilibrium and the core does show that prices are implicitly contained in the idea of a core. The relationship also underlines the fact that more information than contained in prices is needed to find a general competitive equilibrium.

The discussion thus far has centred on perfect information in a general equilibrium model without uncertainty. Putting uncertainty into the model involves changing the specification of the market structure and the informational flow of the model. It is now necessary to know when the uncertainty is resolved to specify how the market reacts. Moreover, it is also necessary to specify the agent's subjective beliefs about the likelihood of the various states of nature. Although the advent of uncertainty raises many interesting questions about imperfect or incomplete information – for example, moral hazard problems when actions are unobservable or adverse selection problems when information is unobservable – questions remain about perfect information in models with uncertainty. In particular, consider an Arrow–Debreu world under uncertainty. In this model the information requirements are analogous to the requirements in a general equilibrium model under certainty with perfect information. In this economy trading takes place for contingent claims or Arrow–Debreu commodities. More precisely, since each state of the world can be distinguished, trading for commodities occurs for each commodity for each state of the world. This increases considerably the number of markets and the number of trades. However, except for information about which state of the world has occurred there are no extra informational requirements in this model. Each agent, knowing his own tastes and endowments in each state of the world, must know only prices. To actually find equilibrium prices, however, excess demands must be known in each possible state of the world.

Perhaps a more reasonable economy under uncertainty is to allow trading to take place on the basis of expectations or beliefs about the likelihood of the states of the world and not to assume that the state of the world is known after trading occurs, that is, not to allow contingent trades. The informational requirement in this model is quite different than in the Arrow–Debreu model. In this model there is only one market clearing price for each commodity, rather, as in the Arrow–Debreu world, than a price for each commodity in each state of the world. The agents (or auctioneer) need not know which state of the world actually occurred. However, they must know which states are possible. Finally, the equilibrium in this model depends crucially on the subjective beliefs of the agents, whereas in the Arrow–Debreu model subjective beliefs do not affect the equilibrium outcomes.

This difference in market structure and information requirement in these two models leads to a loss in efficiency. In the Arrow–Debreu model equilibrium is always Pareto optimal but in the non-contingent claims model it will, in general, not be Pareto optimal. Non-contingent claims equilibrium will in general be *ex ante* but not ex *post* Pareto optimal. In fact, if the market were to reopen after the realization of the state of the world and trading were allowed to take place, a Pareto optimal Arrow–Debreu equilibrium would result.

See Also

▶ Uncertainty

Perfectly and Imperfectly Competitive Markets

John Roberts

In the competition between economic models, the theory of perfect competition holds a dominant market share: no set of ideas is so widely and successfully used by economists as is the logic of perfectly competitive markets. Correspondingly, all other market models (collectively labelled ‘imperfectly competitive’ and including monopoly, monopolistic competition, dominant-firm price leadership, bilateral monopoly and other situations of bargaining, and all the varieties of oligopoly theory) are little more than fringe competitors.

Although it is not surprising that perfect competition should play a central role as a benchmark for normative purposes, the dominance of perfectly competitive forms of analysis in descriptive and predictive work is remarkable. First, economic theorists seem to be increasingly of the view that something like imperfect competition is the fundamental idea, in that perfect competition should be justified by deriving it from models where imperfectly competitive behaviour is allowed and, in particular, agents recognize the full strategic options open to them and any monopoly power they have. This view has led to a large volume of work over the last twenty-five years that, for the most part, suggests that perfect competition corresponds to an extremely special, limiting case of a more general theory of markets. Second, as the idea of perfect competition has been made more precise and the conditions supporting it have become better understood, it has become completely evident that no important market fully satisfies the conditions of perfect competition and that most would not appear even to come close. This is not to say that models should be descriptively accurate; the only way a map could approach descriptive accuracy would be for it to have a scale of 1:1, but such a map is useless. Still, it is striking that economists so consistently opt for a mode with so little apparent descriptive value. Third, the received theory of perfect competition is a theory of price competition that contains no coherent explanation of price formation. That such a fundamental incompleteness does not severely limit the value of the theory is striking.

Given all this, the dominance of perfectly competitive methods should probably be viewed as a reflection of the weakness of imperfectly competitive analysis. There is in fact no powerful general theory of imperfect competition. Instead, there is a myriad of competing partial equilibrium models of imperfectly competitive markets, and the only general equilibrium theories either rely on questionable assumptions or embody institutional specifications that are no more satisfactory than those associated with perfectly competitive analysis.

Despite the unsatisfactory state of both perfectly and imperfectly competitive market theory, recent work based on game-theoretic methodology holds promise of providing a more satisfactory theory of imperfectly competitive markets, of yielding better insight into why perfectly competitive analysis seems to work so well, and of unifying these theories.

Perfect Competition

The idea of perfect competition has many aspects: absence of monopoly power; demand and supply curves that, to the individual, appear horizontal; negligibility of an individual's quantities relative to aggregates; price-taking behaviour (with respect to publicly quoted prices); zero profits and equality of returns across all activities; prices equalling marginal costs and factor returns equalling the values of marginal products; and Pareto-efficiency of market allocations and the efficacy of the Invisible Hand. Stigler (1957) has traced the historical development of the idea of perfect competition essentially through the 'imperfect competition revolution' of the 1930s, noting the appearance of many of these features and documenting the increasing recognition of the stringency of the conditions that appeared to be necessary and/or sufficient for perfect competition. Together these include: large numbers; free entry and exit; full information and negligible search costs; product homogeneity and divisibility; lack of collusion; and absence of externalities and of increasing returns to scale.

The theory about which Stigler wrote still largely corresponds to what is presented in intermediate textbooks and probably to the way most economists think about perfect competition when doing applied work. Firms and consumers are treated as making quantity choices at given prices, because with large numbers, it is suggested, individual quantities are) 'negligible' relative to the aggregate, upon which prices are assumed to depend. (These arguments derive from Cournot 1838.) But how prices are determined is not modelled. This approach is justified by informal arguments that prices are actually set by individual agents, but that, with many agents on each side of the market, any individual would be unable to deviate significantly from the prices charged by others without losing all demand or being overwhelmed by buyers. This idea is connected to the work of Bertrand (1883), but is not supported by formal arguments showing that the outcome of such price setting would be perfectly competitive under the assumed structural conditions (large numbers, homogeneity, free entry, etc.).

When Stigler wrote, Arrow, Debreu and MacKenzie had already provided their path-breaking formal analyses of Walrasian general equilibrium, and within two years Debreu published *Theory of Value* (1959), which is still the standard treatment of this subject. In this theory, competition is given a behavioural definition. There is a given list of consumers and of firms and a given list of commodities. A single price for each good is introduced, and perfectly competitive behaviour is then defined. It involves each consumer selecting the net transactions that maximize utility, subject to a budget constraint defined under the assumptions that the consumer can buy or sell unlimited quantities at the specified prices and that the consumer's purchases do not influence the profits he/she receives. As well, each firm selects the inputs and outputs that maximize its net receipts, again given that the firm can buy and sell any quantities it might consider without influencing prices. Finally, equilibrium is a price vector and perfectly competitive choices for each agent at these prices aggregate to a feasible allocation, that is, such that markets clear.

Three fundamental results are proved for this model. These give conditions on tastes, endowments, and technology under which competitive equilibria exist (existence), equilibrium allocations are Pareto-optimal (efficiency), and, with an initial reallocation of resources, any Pareto optimum can be supported as a competitive equilibrium (unbiasedness). The efficiency and existence theorems together formalize Adam Smith's argument of the invisible hand leading self-interested behaviour to serve the common good, while the unbiasedness result indicates that the competitive price system does not inherently favour any group (capitalists, workers, resource owners, consumers, etc.). The non-wastefulness result requires few assumptions beyond those built into the structure of the model: it is enough that not all consumers are satiated. The existence theorem, however, involves much stricter conditions, including especially the absence of any increasing returns to scale. (This is also needed for the unbiasedness result.)

Many of the conditions arising in less formal treatments of perfect competition are embodied in Debreu's formulation. For example, the very definition of a commodity involves homogeneity, and divisibility is explicitly assumed. Strikingly, however, free entry and large numbers play no explicit role in this theory: all the theorems would hold if there were but a single potential buyer and seller of any commodity.

This numbers-independence property relies crucially on the theory being only an *equilibrium theory*, that is, one which specifies what happens only if behaviour is exactly as stipulated and prices are set at equilibrium, market-clearing values. No examination is offered of what would happen if prices were not at their Walrasian levels, nor indeed, of how prices are determined. Further, not even the famous story of a disinterested Walrasian auctioneer and *tâtonnement* (no trade at nonequilibrium prices) supports this equilibrium by giving a consistent model of price formation with rational actors. Instead there would be incentives to misrepresent demands, responding consistently to each price announcement by the auctioneer as if one had different preferences than actually obtain, with the object of effecting monopolistic prices and outcomes (Hurwicz 1972).

The ability of an individual to manipulate price formation by an auctioneer does disappear once one moves to a model where individuals truly are negligible. Such a model was first introduced by Aumann (1964), where the set of agents is indexed by a continuum endowed with a non-atomic measure. This measure is interpreted as giving the size of a group of agents in comparison with the whole economy. The absence of mass points implies that no individual's excess demands represent a positive fraction of the totals. Thus, any individual's withholding of supply affects neither the magnitude of excess demand (as measured on a per capita basis) nor, correspondingly, whether particular prices clear markets. Thus price-taking is fully rational if prices can be considered to be set by a disinterested auctioneer.

The infinite economy framework captures the large numbers, negligibility, and (with an auctioneer) price-taking aspects of perfect competition. Infinite models also provide a setting where numerous other models of production and exchange agree with the Walrasian in their outcomes. However, infinite models clearly are an extreme abstraction, and the real issue is the extent to which they approximate finite economies. This question leads to consideration of sequences of increasingly large finite economies in which each individual becomes relatively small, perhaps with many others like him or her being present. The identification of perfect competition with such sequences of economies and the asymptotic properties of their allocations dates back to Cournot (1838) and Edgeworth (1881) and has become the basis of several major lines of research.

The most complete of these shows that the core converges to the Walrasian allocations (see Hildenbrand 1974). However, recently attention has focused on the programme initiated by

Cournot of obtaining perfect competition as the limit of imperfectly competitive behaviour and outcomes (see Mas-Colell 1982).

There are three approaches to this problem. One, represented by Roberts and Postlewaite (1976), effectively takes some version of the auctioneer story as given and examines the incentives to respond to price announcements using one's true demands. Here it is shown that if the economy grows through replication or if the sequence of economies under consideration converges to one at which the Walrasian price is locally a continuous function of the data of the economy, then correct revelation of preferences and price-taking is asymptotically a dominant strategy. The second line of work builds more directly on Cournot's model. Agents select quantities and prices somehow arise to clear markets, with some agents (usually the firms) recognizing the impact of their choices on prices and others (consumers) taking prices as given. The central results here are due to Novshek and Sonnenschein (1978), who showed that the freeentry Cournot equilibria converge to the Walrasian allocations as the minimum efficient scale becomes small, provided that a condition of downward sloping demand is met. Finally, the game-theoretic models of noncooperative exchange initiated by Shubik (1973) also lead asymptotically to Walrasian equilibria (see Postlewaite and Schmeidler 1978). A significant feature of these game-theoretic models is that they explicitly treat out-of-equilibrium behaviour: the outcome of *any* pattern of behaviour is specified, not just what happens in equilibrium. This is an important advance. However, in these models, prices appear only as the ratio of the amount of money bid for a good to the amount of the good offered, and are not directly chosen by agents.

A complementary approach to perfect competition (Ostroy 1980) relates to marginal productivity theory and to horizontal demands. Central to this approach is a non-surplus condition that, agent by agent, the rest of the economy would be no worse off if the agent's resources and productive capability were removed from the economy. No-surplus allocations correspond to the economy's having Walrasian equilibria at the same prices with or without any single agent (so demands are horizontal). An economy is defined as perfectly competitive if the no-surplus condition is met. This can happen with a finite number of agents, but typically it requires an infinity.

Thus, various pieces of formal theory capture most of the aspects of the intuitive notion of perfect competition, but this theory points to perfect competition being a limiting case associated many agents in each market or existence of close substitutes for each firm's output, as well as with properties of continuity of the Walras correspondence and downward sloping demand. Also, this theory lacks models in which prices are explicitly chosen by economic agents. None of these results gives much reason for the success that economists have using perfectly competitive analysis.

Imperfect Competition

Formal modelling of markets begins with Cournot's (1838) treatment of quantity-setting, noncollusive oligopoly. Cournot's model yields prices in excess of marginal cost, with this divergence decreasing asymptotically to zero as the number of firms increases. The nineteenth century saw two other important contributions to imperfect competition theory: Bertrand's (1883) pricesetting model which, with constant costs, yields perfectly competitive outcomes from duopoly, and Edgeworth's (1897) demonstration that introducing capacity constraints into this model could prevent existence of (pure strategy) equilibrium.

Thus, even before the important competition revolution, the theory of imperfectly competitive markets was subject to one of the standard complaints still made against it: that it consists of too many models that yield conflicting predictions. This complaint intensified with the proliferation in the 1930s and later of models of firms facing downward-sloping demands. These models usually capture some element of actual competition (or at least appear more realistic than the perfectly competitive alternative). However, it sometimes seems that one can concoct an imperfect competition model that predicts any particular outcome one might wish.

A second complaint against imperfectly competitive analysis is its lack of a satisfactory multiple market formulation.

The first significant contribution to a general equilibrium theory of imperfect competition was Negishi's (1961) model, with later contributions from numerous authors during the 1970s. Although these models differ on important dimensions, the basic pattern in this work involves supplementing the Arrow–Debreu multi-market model of an economy by allowing that some exogeneously specified set of firms perceive an ability to influence prices. (These firms may or may not perceive the actual demand relations correctly.) Equilibrium is then a set of choices (prices or quantities) for each imperfect competitor that maximizes its perceived profits, given the behaviour of the other imperfect competitors and the pattern of adjustment of the competitive sectors (under Walrasian, price-taking behaviour) to the choices of the imperfect competitors.

This theory, as it stood in the mid–1970s, was obviously incomplete on several grounds. Most fundamentally, there was no explanation of why some agents should take prices as given while other agents, who formally might be identical to the price-takers, behave as imperfect competitors. Moreover, it then emerged that there were serious flaws in the crucial existence theorems that purported to show that the models were not vacuous.

These theorems obtained profit maximizing choices for the imperfect competitors that were mutually consistent by use of fixed-point arguments based on Brouwer's theorem. To use these methods, the optimal choices of any one agent must depend continuously on the conjectured choices of the others. This role of continuity of reaction functions is analogous to that of continuity of demand functions in the Arrow–Debreu model. However, unlike the continuity of demand, continuity of reaction functions was not derived from conditions on the fundamental data of the economy. Rather, it was either directly assumed or obtained by supposing that the imperfect competitors' perceptions of demand yielded concave profit functions.

Roberts and Sonnenschein (1977) showed that this approach was problematic by displaying extremely simple, nonpathological examples in which reaction functions are discontinuous and no imperfectly competitive equilibrium exists. The source of these failures is nonconcavity of the profit functions, and no standard conditions on preferences ensure the needed concavity: it can fail with only a single consumer or when all consumers have homothetic preferences. (Note, however, that existence ceases to be a problem in general equilibrium Cournot models if the economy, including the number of imperfect competitors, is made large enough through replication.)

These problems with imperfect competition theory perhaps explain some of the popularity of perfect competition models. However, they also suggest two important, positive points. First, the multiplicity of models and the divergence in their predictions indicates that, at least in small numbers situations, institutional details are important. Economists, habituated to the use of perfectly competitive methods, typically are imprecise about such factors as how prices are actually determined, whether decisions are made simultaneously or sequentially, whether individuals select prices, quantities, or both, and what happens when agents' plans are inconsistent. These factors cannot be treated so cavalierly in dealing with imperfectly competitive models and probably ought not to be when actual markets are being analysed. Second, both the failure of existence in models of imperfectly competitive general equilibrium and the unexplained asymmetry of assumed behaviour in these models suggest that a simple grafting of imperfect competitors onto the standard Arrow–Debreu model will not yield a satisfactory theory. Rather, one ought to start afresh from the foundations with a more careful modelling.

Strategic Models of Competition

An approach to both of these points is provided by the methods of the theory of noncooperative games and especially games in extensive form. Recent work using this approach has resulted in significant improvements in the partial equilibrium theory of imperfect competition, and there is reason to hope that these same methods can provide a satisfactory general equilibrium theory. Moreover, this approach also offers hope of ultimately yielding a unified theory of competition that would encompass both perfect and imperfect competition.

To model a market as a game in extensive form, one must specify the set of participants, the beliefs each has about the characteristics of the other agents, the order in which each acts, the information available to each whenever it makes a decision, the possible actions available at each decision point, the physical outcomes resulting from each possible combination of choices, and the valuations of these outcomes by the agents. Thus, such a model involves a complete specification of a particular set of institutions. This aspect might be viewed as a drawback, but it is in fact a potential strength of these methods.

(Note that adopting this approach does not require that price formation be modelled by having prices be chosen by agents in the model. Indeed, Cournot's original model is a well-specified game, but price formation is not explicitly modelled. However, this framework does facilitate and encourage such a specification.)

Given a game, one next specifies a solution concept. In principle, there is great freedom in making this specification, but most researchers opt for the Nash equilibrium or some refinement thereof. Note that adopting the Nash equilibrium does not rule out collusion if opportunities to coordinate and to enforce agreements are modelled as part of the game. Nor does it mean that the agents are acting simultaneously: the order of moves is part of the specification of the game, and the Nash equilibrium applies equally to simultaneous or sequential moves. To illustrate, the von Stackelberg solution corresponds to subgame-perfect Nash equilibrium in a game where the designated leader moves first and the follower observes the leader's choice before making its own. Finally, the Nash criterion does not restrict analysis to one-shot situations; it is equally applicable to models of repeated play.

When von Neumann and Morgenstern's (1944) treatise on game theory first appeared, there was hope among economists that these methods would unify and advance the analysis of imperfect competition. When these hopes were not quickly realized, many economists wrote off game theory as a failure. This position is still reflected in many intermediate textbooks. However, in the last decade these hopes have been revitalized by actual accomplishments of these methods.

The first contribution of this work has been to begin unifying the existing theory of imperfect competition. This has been done on one level by providing a common language and analytical framework in terms of which earlier work can be cast and understood. In this line, game theoretic treatments have made formal sense out of such ideas as reaction curves and kinked demand curves by obtaining equilibria of well-specified, dynamic games that have these features. As well, various of the older theories that appeared to be in conflict have been shown to be consistent in that they arise from a common, more basic model. For example, the Cournot and the von Stackelberg solutions can both be attained as Nash equilibria in a single model where the timing of moves is endogenous. In a similar vein, the Cournot, Bertrand and Edgeworth models have been integrated by showing that equilibrium in a two-stage game where duopolists first select capacities and then compete on price yields the Cournot quantities.

A second contribution has been to provide models embodying aspects of imperfect competition that had been widely discussed in the industrial organization literature but previously lacked formal expression. The best example here is work showing how limit pricing, predatory pricing, and price wars can arise as rational behaviour in the presence of informational asymmetries between competitors (see Roberts 1986). Further examples include explanations of sales and other discriminatory pricing policies, the determination and maintenance of product quality, the use of capacity and other investments in commitment to deter entry, and the opportunities for and limitations on implicit collusion. This work is revolutionizing the field of industrial organization.

The third contribution has been to permit the analysis of realistic models of institutions for exchange actually present in the economy. The best-developed example of such work is that on auctions to sell a single object to one of many potential buyers (see Milgrom 1986), but important work has also been done on multi-object auctions and other monopoly pricing institutions (including posted prices, priority pricing, and nonlinear pricing), bilateral monopoly and bargaining, and bid-ask markets or oral double auctions. In this work, the rules of the institution being modelled, the distribution of information about tastes, costs, etc., held by the various participants, and the preferences of these agents together induce a game in extensive form. This game captures the full strategic options open to all the participants, specifying completely the prices and allocations resulting from any choice of actions. Thus, the Nash equilibrium of this game yields explicit predictions of the choices of prices and of the volume, timing, and pattern of trade. Often these predictions are both remarkably tight and in agreement with observed behaviour.

This work is providing a more complete description and a clearer theoretical understanding of the operation of actual markets. Moreover, by providing detailed predictions of the outcomes of equilibrium behaviour under different institutions, it gives the basis for a theory of the choice among market institutions (see, for example, Harris and Raviv 1981). Finally, it provides an approach to unifying the theories of perfect and imperfect markets and market behaviour. In this work, agents' behaviour is rationally strategic relative to the given economic situation. However, in particular environments this imperfectly competitive behaviour may be very close to perfectly competitive or may yield outcomes that are essentially competitive (see Wilson 1986). By determining the situations in which this is true, we may finally understand when and why perfectly competitive analyses succeed.

See Also

- ▶ Competition
- ▶ Imperfect Competition
- ▶ Monopolistic Competition and General Equilibrium
- ▶ Nash Equilibrium

Bibliography

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.

Bertrand, J. 1883. Théorie mathématique de la richesse sociale. *Journal des Savants* 48: 499–508.

Cournot, A. 1838. *Recherches sur les principes mathématiques de la théorie des richesses*. Paris: M. Rivière.

Debreu, G. 1959. *The theory of value*. New York: Wiley.

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

Edgeworth, F.Y. 1897. La teoria pura del monopolio. *Giornale degli Economisti* 15: 13ff.

Harris, M., and A. Raviv. 1981. A theory of monopoly pricing schemes with demand uncertainty. *American Economic Review* 71: 347–365.

Hildenbrand, W. 1974. *Core and equilibria of a large economy*. Princeton: Princeton University Press.

Hurwicz, L. 1972. On informationally decentralized systems. In *Decision and organization*, ed. C.B. McGuire and R. Radner, 297–336. Amsterdam: North-Holland.

Kalai, E., and W. Stanford. 1985. Conjectural variations strategies in accelerated Cournot games. *International Journal of Industrial Organization* 3: 133–152.

Kreps, D.M., and J.A. Scheinkman. 1983. Quantity precommitment and Bertrand competition yield Cournot outcomes. *Bell Journal of Economics* 14: 326–337.

Mas-Colell, A. (ed.). 1982. *Non-cooperative approaches to the theory of perfect competition*. New York: Academic.

Milgrom, P.R. 1986. Auction theory. In *Advances in economic theory*, ed. T. Bewley. Cambridge: Cambridge University Press for the Econometric Society.

Negishi, T. 1961. Monopolistic competition and general equilibrium. *Review of Economic Studies* 28: 196–201.

Novshek, W., and H. Sonnenschein. 1978. Cournot and Walras equilibrium. *Journal of Economic Theory* 19: 223–266.

Ostroy, J. 1980. The no-surplus condition as a characterization of perfectly competitive equilibrium. *Journal of Economic Theory* 22: 183–207.

Postlewaite, A., and D. Schmeidler. 1978. Approximate efficiency of non-Walrasian equilibria. *Econometrica* 46: 127–137.

Roberts, J. 1986. Battles for market share: Incomplete information, aggressive strategic pricing, and competitive dynamics. In *Advances in economic theory*, ed. T. Bewley. Cambridge: Cambridge University Press for the Econometric Society.

Roberts, J., and A. Postlewaite. 1976. The incentives for price-taking behavior in large exchange economies. *Econometrica* 44: 115–127.

Roberts, J., and H. Sonnenschein. 1977. On the foundations of the theory of monopolistic competition. *Econometrica* 45: 101–113.

Shubik, M. 1973. Commodity money, oligopoly, credit and bankruptcy in a general equilibrium model. *Western Economic Journal* 11: 24–38.

Stigler, G. 1957. Perfect competition, historically contemplated. *Journal of Political Economy* 65: 1–17.
von Neumann, J., and O. Morgenstern. 1944. *Theory of games and economic behavior*. Princeton: Princeton University Press.
Wilson, R. 1986. Game theoretic analyses of trading process. In *Advances in economic theory*, ed. T. Bewley. Cambridge: Cambridge University Press for the Econometric Society.

Performing Arts

William J. Baumol

In the past two decades a substantial international literature on the economics of the arts has accumulated. Aside from the importance of the cultural contribution made by the arts, interest in the subject among economists has been elicited by some special attributes of the economics of the arts which have proved interesting analytically and whose analysis has had significant applications outside the field. Notable is the 'cost disease of the performing arts' which has been proposed as an explanation for the fact that, except in periods of rapid inflation, the costs of artistic activities almost universally rise (cumulatively) faster than any index of the general price level. Another major theoretical issue with which the literature has concerned itself is the grounds on which public sector funding of the arts can be justified.

Organization and Funding

The structure of the performance industry is similar in many of the industrialized countries. The largest enterprise in terms of budget and personnel is the opera, followed, in rank order, by the orchestra, theatre and dance. The theatres are the only group that contains a substantial profit seeking sector. All of the others, and many of the theatres as well, receive a substantial share of their incomes from government support and private philanthropy. The US, with its policy of tax exemptions, is probably the only country in which the share of private philanthropy is large, and there it exceeds the amount of government funding by a large margin. In many countries the bulk of such financing is provided by only a single agency, while in the US an arts organization whose application has been rejected by one funding source can usually turn to others for reconsideration.

The available statistical evidence suggests that demand for attendance is fairly income elastic but quite price inelastic, at least in the long run. This suggests that the widely espoused goal of diversity in audiences prevents ticket prices from rising more than they have, although fear that such rises will cause temporary but substantial declines in revenues and will reduce philanthropic or government support no doubt also plays a part.

In every country in which systematic audience studies have been carried out, the audience has been shown to be drawn from a very narrow range. It is far better educated than the average of the population, it has a far higher average income, it is somewhat older, and it includes a remarkably small proportion of blue-collar workers. Even free or highly subsidized performances affect this only marginally.

While total expenditures on ticket purchases have, of course, risen over the years, the pattern is modified substantially when corrected for changes in population, the price level and real incomes. Thus, in the US, the share of per capita disposable income devoted to admissions to artistic performances declined from about $0.15 out of every $100 in 1929 to about $0.05 in 1982. The latter figure has been virtually unchanged throughout the period since World War II.

The Cost Disease of the Performing Arts

One of the special features of the economics of the performing arts that seems to colour their cost structure is their 'cost disease'. This condemns the cost of live performance to rise at a rate persistently faster than that of a typical manufactured good. An illustration comparing the costs of watchmaking and of musical performance over the centuries shows the reason. There has been vast and continuing technical progress in watchmaking, but live performance benefits from no labour-saving innovations – it is still done the old-fashioned way. Toward the end of the 17th century a Swiss craftsman could produce about 12 watches per year. Three centuries later that same amount of labour produces over 1200 (non-quartz) watches. But a piece of music written three centuries ago by Purcell or Scarlatti takes exactly as many person hours to perform today as it did in 1685 and uses as much equipment.

These figures mean that while one has to work just about as many hours to pay for a ticket to an opera today as one would have in similar jobs 300 years ago, the cost of a watch or of any other manufactured good has plummeted, in terms of the labour time we must pay for it. In other words, because manufactured goods have benefited from technological advance year after year while live performances have not, almost every year theatre and concert tickets have become more and more expensive in comparison with the price of watches. This phenomenon has been called 'the cost disease of live performance'.

To facilitate comparison with the discussion of the cost structure of the mass media that follows, it is helpful to describe the cost disease formally. Let

y_{it} = output of product *i* in period *t*
x_{kit} = quantity of input *k* used in producing *i*
AC_{it} = average cost of *i* in period *t*
w_{kt} = (real) price of *k* in period *t*
$\pi_{it} = y_i / \sum w_{kt}\, x_{kit}$ = total factor productivity in output *i*
* = rate of growth, i.e. for any function, $f(t)$
$f^* = \cdot f/f.$

Then we have:

Proposition 1 Let y_{1t} and y_{2t} be two outputs produced by single product firms. Then, if $\pi^*_{1t} \leq r_1 < r_2 \leq \pi^*_{2t}$, so that output 1 may be called relatively 'stagnant' (and output 2 is relatively) 'progressive'), the ratio of the average cost of output 1 to that of output 2, AC_{1t}/AC_{2t} will rise without limit.

Proof By definition

$$AC_{1t}/AC_{2t} = \pi_{2t}/\pi_{1t}$$

so that

$$(AC_{1t}/AC_{2t})^* = \pi^*_{2t} - \pi^*_{1t} \geqslant r_2 - r_1.\ \text{Q.E.D.}$$

Here, of course, y_1 may be interpreted as the output of live performance and y_2 as the output of manufactured goods. It follows that the prices of manufactured goods can be expected to rise less quickly than those of concerts, dance or theatrical performances. Ticket prices must therefore rise faster than the economy's overall rate of inflation, since the latter is an average of the increases in the prices of all the economy's goods.

It is sometimes suggested that the mass media – film, radio, television and recording – can provide the cure for the cost disease, but recent analysis suggests that despite their sophisticated technology many of the mass media are in the long run vulnerable to essentially the same problem. As a matter of fact, the data indicate that the cost of cinema tickets and the cost per prime-time television hour have been rising at least as fast as the price of tickets to the commercial theatre. The explanation apparently lies in the structure of mass media production, which is made up of two basic components that are very different technologically. The first comprises preparation of material and the actual performance in front of the cameras, while the second is the transmission or filming.

Television broadcasting of new material requires these two elements in relatively fixed physical proportions – one hour of programming (with some flexibility in rehearsal time) must be accompanied by one hour of transmission for every one hour broadcast. However, since the first component of television is virtually identical with live performance on a theatre stage, there is just as little scope for technical change in the one as in the other, while the second component, on the other hand, is electronic and 'high tech' in character and constantly benefits from innovation.

Industries with this cost structure have been referred to as) ‘asymptotically stagnant’. The evolution of such an industry over time is characterized by an initial period of decline in total cost (in constant dollars) which *must* be followed by a period in which its costs begin to behave in a manner more and more similar to the live performing arts. The reason is that the cost of the highly technological component (transmission cost) will decline, or at least not rise as fast as the economy’s inflation rate. At the same time, the cost of programming increases at a rate surpassing the rate of inflation.

If each year transmission costs decrease and programming expenses increase because of the cost disease that besets all live performance, eventually programming cost must begin to dominate the overall budget. Thereafter, total cost and programming cost must move closer and closer together until virtually the entire budget becomes a victim of the disease, with the stable technological costs too small a fragment of the whole to make a discernible difference.

These results are encompassed in the following propositions:

Proposition 2 Suppose an activity, A, uses stagnant input x_1 and progressive input x_2 in fixed proportion v, so that $x_{2t} = vx_{1t}$. If w_{1t}, the unit price of x_{1t}, increases at a nonnegative rate no less than r_1 and w_{2t} increases at a rate no greater than r_2, where $r_2 < r_1$, then the share of total expenditure by A that is devoted to x_{1t} will approach the limit unity. Moreover, for any g such that $0 < g < 1$, there exists T such that for all $t > T$

$$1 \geqslant w_{1t}x_{1t}/(w_{1t}x_{1t} + w_{2t}x_{2t}) \geqslant 1 - g.$$

Proof We are given

$$\begin{aligned} &w_{1t} \geqslant a_1 \mathrm{e}^{r_1 t} \\ &w_{2t} \leqslant a_2 \mathrm{e}^{r_2 t}, x_{2t} = vx_{1t}. \end{aligned}$$

Then,

$$1 \leqslant \frac{w_{1t}x_{1t} + w_{2t}x_{2t}}{w_{1t}x_{1t}} = 1 + \frac{\upsilon w_{2t}}{w_{1t}} \leqslant 1 + (a_2/a_1)\upsilon\,\mathrm{e}^{(r_2 - r_1)^t}\cdot \text{ Q.E.D}$$

Along similar lines one can prove:

Proposition 3 Let A in Proposition 2 be supplied under conditions of perfect competition, and let its output, y_t satisfy $y_1 = ux_{1t}$ (u constant) and let its price be p_t. The p^* will approach that of the price of its stagnant input.

Corollary The smaller the value of w_{2t}, i.e. the more progressive is the progressive input of A, the more rapidly will the behaviour of A’s price approximate to that of its stagnant input.

Grounds for Public Support

Several economists have explored the grounds, if any, on which public support for the performing arts can be justified. They have examined all the usual criteria and found most of them weak. For example, income distribution concerns surely do not explain public financing of activities consumed largely by persons with incomes above the average. The beneficial externalities of attendance of the arts are not only difficult to document but are even hard to describe in the abstract. The same is true of the public good properties of performance. The best that has been done is to argue (1) that they have an ‘option value’ – even those who do not care to attend, themselves, may want to keep the arts alive for their grandchildren; and (2) that they constitute a partially public good through their part in the educational process and the (national) pride they engender even in those who do not attend themselves (or the embarrassment they avoid among those who do not want to belong to a nation of philistines). In the last analysis, it is simply argued that the arts deserve support because they are) ‘merit goods’ (to use Musgrave’s term). But that amounts to substitution of nomenclature for analysis. What the discussion comes down to is that the evidence suggests strongly that the public considers the arts worth supporting, and that in a democracy the public has the right to support what it wants to. Welfare theory has little to contribute here.

The cost disease analysis has been used by administrators throughout the world as justification for support but, of course, the fact that an activity is under financial pressure is, by itself, no valid reason for public subvention, as economic theory shows so clearly. However, if support is decided upon on other grounds, the cost disease analysis does legitimately help to give guidance on the amounts it will be appropriate to provide. It also warns us of the dangers of underfinancing as a result of what W.E. Oates has called 'fiscal illusion'. The cost disease implies that the cost of performance will rise faster than the general price level. If so, when government support for the arts increases only marginally faster than the general price level, politicians are likely to conclude that, though they have increased the real level of support, the quantity and quality of activity the public is getting for its money is declining. Mismanagement and waste are then likely to be blamed and budgets may be trimmed, on those grounds, below the level that is called for by the public's actual preferences.

Bibliography

Baumol, H., and W.J. Baumol (eds.). 1984. *Inflation and the performing arts*. New York: New York University Press.

Baumol, W.J., and W.G. Bowen. 1966. *Performing arts: The economic dilemma*. New York: Twentieth Century Fund.

Blaug, M. (ed.). 1976. *The economics of the arts*. London: Martin Robertson.

Feld, A.L., M. O'Hare, and J.M.D. Schuster. 1983. *Patrons despite themselves: Taxpayers and arts policy*. New York: New York University Press.

Netzer, D. 1978. *The subsidized muse*. New York: Cambridge University Press.

Throsby, C.D., and G.A. Withers. 1979. *The economics of the performing arts*. New York: St Martin's Press.

Period of Production

G. O. Orosel

Keywords

Aggregate capital; Böhm-Bawerk, E. von; Capital intensity; Capital measurement; Capital theory; Non-substitution theorem; Period of production; Rate of profit; Reswitching; Stationary state; Time preference

JEL Classifications

D2

The period of production is, or purports to be, a measure of aggregate capital per head. More specifically, it is a theoretical concept which tries to measure an economy's (heterogeneous) capital stock per head in (homogeneous) units of time.

Necessarily the concept of the period of production is based on an Austrian, or temporal, view of production. In this view production is conceptualized as a sequence of primary inputs on the one hand and a corresponding sequence of consumption outputs on the other. Produced means of production (capital goods) are reduced to dated primary inputs and consumption outputs. This implies that the approach is suited best to the analysis of steady states, where specific properties of capital goods are irrelevant, whereas it will be misleading, in general, if applied to problems of transition or disequilibrium. In particular, this approach is inadequate for business cycle analysis.

Although the temporal view can be traced back to Thünen, Senior, Rae and Jevons, it was Böhm-Bawerk (1889) who made it a cornerstone of his theory. This theory was directed at a fundamental problem of political economy: why is the (net) rate of profit positive? A related problem concerns the measurement of heterogeneous capital goods in homogeneous units which are independent of distribution.

A sketch of Böhm-Bawerk's theory is as follows. According to Böhm-Bawerk the fundamental feature of an economy using capital is that there is a temporal distance, called *period of production* (or *period of investment*), between primary inputs and corresponding consumption outputs. Capital is, in its essence, a fund of means of subsistence which allows for consumption during this period. In a steady state this subsistence fund consists of different 'layers' of goods which are distinguished by their respective degree of maturity, such that each period's consumption can be provided by the layer which has just become ready for consumption. A longer period of production is equivalent, in this view, to more capital per head. Hence the per capita stock of heterogeneous capital goods can be measured in homogeneous units of time. Adding to this (*a*) the technological hypothesis that consumption output per head increases with the period of production, and (*b*) the psychological hypothesis of a positive time preference gives, in a nutshell, Böhm-Bawerk's explanation of the positivity of the rate of profit.

From the beginning, Böhm-Bawerk's theory, and in particular the concept of the period of production, has caused heated debates (involving, among others, J.B. Clark, Irving Fisher, Schumpeter, Wicksell, Hayek, Kaldor and Knight). The contributions to these debates, not all of them to the point, are not reviewed here (see, however, Kaldor 1937; Weston 1951). Instead, we will analyse the period of production from a *fundamentalist* and from a pragmatic point of view. In a fundamentalist view the period of production is seen as an important component of the theory sketched above and, therefore, must have properties which make it consistent with this theory. In particular, it must be a technological parameter. In a *pragmatic* view the period of production is just a conventionally measured distance between primary inputs and consumption outputs and need not have any definite properties.

In order to give a more rigorous presentation of the period of production and the problems associated with it, we make the following assumptions. Unless stated otherwise time is measured continuously and it is assumed that primary inputs and consumption outputs can each be measured in homogeneous units. A technique is assumed to be representable by a pair (*a, b*) of non-negative, continuous functions a: $R \rightarrow R_+$ and b: $R \rightarrow R_+$ where $a(t)$ is the amount of primary inputs expended at *t* and b(*t*) is the amount of consumption outputs delivered at *t* (note that such a representation where (*a, b*) is independent of the rate of growth may not be possible for technologies with joint production; cf. the non-substitution theorem). The primary input will be called 'labour'; 'per head' (or 'per capita') will mean per unit of labour. It is assumed that

$$\lim_{t\to-\infty} a(t) = \lim_{t\to-\infty} a(t) = \lim_{t\to-\infty} b(t) = \lim_{t\to-\infty} b(t) = 0$$

that *a* and *b* are not identically zero, that there are constant returns to scale (that is, for any feasible technique (*a, b*) and any $\lambda > 0$ the technique (λa, λb) is also feasible) and that there exist some real numbers $H < 0,\ G > 0$ such that the improper Riemann-integrals

$$\int_{-\infty}^{\infty} \mathrm{e}^{-\gamma t} a(t)\mathrm{d}t$$

and

$$\int_{-\infty}^{\infty} \mathrm{e}^{-\gamma t} b(t)\mathrm{d}t$$

converge for $\gamma \in (H,G)$. The analysis is restricted to steady states with technique (*a, b*), a rate of growth g $\in$ *(H,G)* and a rate of interest $r \in (H,G)$ and to conditions of zero excess profits, implying

$$w\int_{-\infty}^{\infty} \mathrm{e}^{-rt} a(t)\mathrm{d}t = \int_{-\infty}^{\infty} \mathrm{e}^{-rt} \mathrm{b}(t)\mathrm{d}t \qquad (1)$$

where w is the steady state price of the primary input, henceforth called (real) wage, and the price of the consumption good is set equal to 1. Given a technique (*a, b*) and any point of time s, let λ (*s, t*) denote the activity level, at s, of the techniques which are in 'stage' *t*. For a steady state λ (*s, t*) = $\mathrm{e}^{-\mathrm{gt}}$ λ (*s*, 0) and total labour inputs at s are

$$A(s) = \int_{-\infty}^{\infty} \lambda(s,t)a(t)\mathrm{d}t = \lambda(s,0)\int_{-\infty}^{\infty} \mathrm{e}^{-gt} a(t)\mathrm{d}t.$$

Similarly, total consumption outputs at *s* are

$$B(s) = \lambda(s,0)\int_{-\infty}^{\infty} \mathrm{e}^{-gt} b(t)\mathrm{d}t.$$

This implies for per capita consumption *c*: = *B*(*s*)/*A*(*s*)

$$c\int_{-\infty}^{\infty} \mathrm{e}^{-gt}a(t)\mathrm{d}t = \int_{-\infty}^{\infty} \mathrm{e}^{-gt}b(t)\mathrm{d}t \qquad (2)$$

which is dual to (1). For the value of capital k per head the steady state identity.

$c + gk = w + rk$ implies

$$k = \begin{cases} \dfrac{c-w}{r-g} & \text{for } r \neq g \\ -\dfrac{\mathrm{d}c}{\mathrm{d}g} = -\dfrac{\mathrm{d}w}{\mathrm{d}r} & \text{for } r = g \end{cases}. \qquad (3)$$

The Fundamentalist View

In a fundamentalist view the period of production *T* must have two properties: first, it must be a technological parameter, that is, for each technique (*a, b*) and rate of growth g it must be uniquely determined by the associated flows of labour inputs and consumption outputs (which are proportional to $\mathrm{e}^{-gt}a(t)$ and $\mathrm{e}^{-gt}b(t)$ respectively); second, as a subsistence fund for *T* periods with per period consumption c the steady state value of capital per head k must be given by $k = cT$. But this leads to an inconsistency: since it implies $T = k/c$ and since in general *k/c* varies with the rate of interest, *T* cannot be a technological parameter. Hence the period of production in the fundamentalist sense does not exist. An analogous inconsistency occurs, if one follows Böhm-Bawerk in (wrongly) identifying consumption with wages and therefore postulates $k = wT$ rather than $k = cT$.

Part of the fundamentalist perspective can be rescued if one gives up the idea that the period of production is one-dimensional. This has been shown by Orosel (1979) within the context of a flow input–point output model where time is measured discretely.

The basic idea can be sketched for a stationary state. With time measured discretely a (flow input–point output) technique can be described by oness consumption output $\beta(0) > 0$ and corresponding labour inputs $\alpha(t) \geq t = 0, -1, -2, \ldots$, where, for some $G>0$,

$$0 < \sum_{t=-\infty}^{0} (1+\gamma)^{-t}\alpha(t) < \infty$$

for $\gamma \in (-1,G)$. To the sequence of labour inputs $\{\alpha(t)\}_{-\infty}^{0}$ is associated a sequence of wage payments $\{z_1(t) = w\alpha(t)\}_{-\infty}^{0}$; a sequence of simple interest payments on these, that is

$$\left\{z_2(\mathrm{t}) = r\sum_{\tau=-\infty}^{t-1} z_1(\tau)\right\}_{-\infty}^{0};$$

a sequence of simple interest payments on z_2(t), that is

$$\left\{z_3(\mathrm{t}) = r\sum_{\tau=-\infty}^{t-1} z_2(\tau)\right\}_{-\infty}^{0},$$

and so on, that is

$$z_{i+1}(t) = r\sum_{\tau=-\infty}^{t-1} z_i(t),$$

i=1,2,... To each sequence $\{z_i(t)\}_{-\infty}^{0}$ we can define a ‘period of production’ *Ti* as the ‘average distance’ of $\{z_i(t)\}_{-\infty}^{0}$ from output $\beta(0)$, that is, from $t = 0$, by

$$T_i := \frac{\sum_{t=-\infty}^{0} (-t)z_i(t)}{\sum_{t=-\infty}^{0} z_i(t)}.$$

Further, with each period of production Ti we can associate a (per capita) subsistence fund s_i which makes it possible to consume the incomes $\{z_i(t)\}_{-\infty}^{0}$ generated by the technique before the technique generates a consumption output. These funds are given by $s_1 = wT_1$ for wages, by $s_2 = (rs_1)T2$ for simple interest on wages, and so on, that is, $s_{i+1} = (rs_i)T_{i+1}$, $i = 1, 2, \ldots$, for simple interest on s_i during the period of production T_{i+1}, associated with these interest incomes. The total per capita subsistence fund is given by

$$s = \sum_{i=1}^{\infty} s_i = wT_1 + \sum_{i=1}^{\infty} (rs_i)\mathrm{T}_{i+1},$$

which is a sum of consumption terms (*w* and *rsi* respectively) each of which is multiplied by the associated period of production. It can be shown that for $r \in [0,G)$ all series converge and (i) all *Ti* are technological parameters; (ii) $k = s$, that is, the value of capital (per head) equals the subsistence fund (per head); (iii)

$$c = w + \sum_{i=1}^{\infty} rs_i,$$

that is, the consumption terms in s add up to per capita consumption. These results can be generalized to steady states with a positive rate of growth (Orosel 1979).

The periods of production *Ti* are fundamentalist in the sense that they are technological parameters and that the subsistence fund corresponding to them equals the value of the capital stock. They lead to a consistent reformulation of some of Böhm-Bawerk's main ideas, but they do not give a measure of aggregate capital. In fact, in the 1960s the debates in the theory of capital have made clear that such a measure does not exist.

The Pragmatic View

There are three prominent proposals as to how to measure the time interval between primary inputs and consumption outputs. They are associated with the names of (i) Böhm-Bawerk (1889), (ii) Hicks (1939) and von Weizsäcker (1971), and (iii) Dorfman (1959). Although only von Weizsäcker's analysis is directly applicable to steady states with a given (flow input–flow output) technique (*a, b*), all three proposals can be generalized accordingly. These three (generalized) concepts of the period of production, denoted by T^B, T^H and T^D respectively, are defined as follows (all integrals being improper Riemann-integrals):

$$T^B(g) := \frac{\int_{-\infty}^{\infty} t\mathrm{e}^{-gt}b(t)\mathrm{d}t}{\int_{-\infty}^{\infty} \mathrm{e}^{-gt}b(t)\mathrm{d}t} - \frac{\int_{-\infty}^{\infty} t\mathrm{e}^{-gt}a(t)\mathrm{d}t}{\int_{-\infty}^{\infty} \mathrm{e}^{-gt}a(t)\mathrm{d}t}, \quad g \in (H,G) \tag{4}$$

$$T^{\mathrm{H}}(r) := \frac{\int_{-\infty}^{\infty} t\mathrm{e}^{-rt}b(t)\mathrm{d}t}{\int_{-\infty}^{\infty} \mathrm{e}^{-rt}b(t)\mathrm{d}t} - \frac{\int_{-\infty}^{\infty} t\mathrm{e}^{-rt}wa(t)\mathrm{d}t}{\int_{-\infty}^{\infty} \mathrm{e}^{-rt}wa(t)\mathrm{d}t}, \quad r \in (H,G) \tag{5}$$

$$T^{\mathrm{D}}(g,r) := \frac{k(g,r)}{c(g)}, g \in (H,G), \ r \in (H,G) \tag{6}$$

where *k*(*g*, *r*)/*c*(*g*) is, in value terms, the capital–consumption ratio (if, as in Dorfman's analysis, a stationary state is considered, it is also the capital–output ratio). Given our assumptions all integrals are convergent. Definitions (4) and (5) measure the difference between two points of gravity, or mean values of time, associated with outputs and inputs respectively (the densities being

$$\mathrm{e}^{-gt}b(t) / \int_{-\infty}^{\infty} \mathrm{e}^{-gt}b(\mathrm{t})$$

and so on). In (4) the densities applied are given by the respective steady state *quantities*, in (5) they are given by the steady state *values*. The justification of (6) is less obvious. Dorfman's argument is that, given g and *r, k* is a constant stock (of value) with a constant outflow *c*; therefore, the average time a unit of *c* remains in *k* is *k*/*c* ('bathtub theorem'). Alternatively, (6) can be derived from the postulate that the (per capita) subsistence fund associated with T^{D}, that is, cT^{D}, equals k.

What are the properties of T^{B}, T^{H} and T^{D}, and how are the three concepts related to each other? First, it is interesting, though not shown in the literature, that T^{D} can also be represented as a difference between points of gravity. Without loss of generality, let the level of activity associated with *t* be $\mathrm{e}^{-\mathrm{gt}}$. Then to a point of time *t* there corresponds a technique ($\mathrm{e}^{-\mathrm{gt}}$ *a*, $\mathrm{e}^{-\mathrm{gt}}$ *b*) and therefore wages $w\mathrm{e}^{-\mathrm{gt}}$ *a*(*t*), profits $r\kappa(t)$ and investments $g\kappa(t)$ where

$$\kappa(t) := \int_{-\infty}^{t} \mathrm{e}^{r(t-\tau)}[w\mathrm{e}^{-gt}a(\tau) - e^{-gt}b(\tau)]\, d\tau$$

is the accumulated value of capital, at t, associated with process ($\mathrm{e}^{-\mathrm{gt}}$ *a*, $\mathrm{e}^{-\mathrm{gt}}$ *b*). Therefore, in a steady state with technique (*a*, *b*), growth rate g and interest rate r there is associated to each t an amount $q(t)$ of consumption claims (wages plus profits minus investments)

$$q(t) := \mathrm{e}^{-gt} wa(t) + (r - g)\kappa(t) \qquad (7)$$

It is possible to prove that these claims sum up to total consumption, that is,

$$\int_{-\infty}^{\infty} q(t)\mathrm{d}t = \int_{-\infty}^{\infty} \mathrm{e}^{-gt} b(t)\mathrm{d}t \qquad (8)$$

and that T^{D} is the 'temporal distance' between consumption outputs and consumption claims, that is

$$T^{\mathrm{D}} = \frac{\int_{-\infty}^{\infty} t\mathrm{e}^{-gt} b(t)\mathrm{d}t}{\int_{-\infty}^{\infty} \mathrm{e}^{-gt} b(t)\mathrm{d}t} - \frac{\int_{-\infty}^{\infty} tq(t)\mathrm{d}t}{\int_{-\infty}^{\infty} q(t)\mathrm{d}t} \qquad (9)$$

In (9) T^{D} has a structure analogous to T^{B} and T^{H}. Because of (4), (5), (7) and (9)

$$T^{\mathrm{B}} = T^{\mathrm{B}} = T^{\mathrm{D}}, \text{for } r = \mathrm{g}. \qquad (10)$$

Differentiation of (2) gives

$$T^{B} = -\frac{1}{c}\frac{\mathrm{d}c}{\mathrm{d}g};$$

of (1)

$$T^{H} = -\frac{1}{w}\frac{\mathrm{d}w}{\mathrm{d}r}.$$

Therefore, using (3), $k = cT^{\mathrm{B}} = wT^{\mathrm{H}} = cT^{\mathrm{D}}$ for $r = g$. For $r \neq g$ we have

$$\frac{1}{r-g}\int_g^r c(\gamma)T^{\mathrm{B}}(\gamma)d\gamma = -\frac{1}{r-g}\int_g^r \frac{\mathrm{d}c(\gamma)}{d\gamma}d\gamma$$
$$= \frac{c(g) - w(r)}{r-g} = k(g, r)$$

since $c(r) = w(r)$ Similarly

$$\frac{1}{r-g}\int_g^r w(\rho)T^{\mathrm{H}}(\rho)\mathrm{d}\rho = k(g, r).$$

Hence k can be interpreted as an average of subsistence funds of the form cT^{B} and wT^{H} respectively. Finally, if for two techniques (a^1, b^1) and (a^2, b^2) one of the three periods of production, $T^{\mathrm{B}}(g)$, $T^{\mathrm{H}}(r)$ or $T^{\mathrm{D}}(g, r)$, is for all feasible g and r greater for (a^1, b^1) than for (a^2, b^2), then for these techniques no reswitching or other paradoxa can occur and (a^1, b^1) can be regarded as unambiguously more capital intensive than (a^2, b^2). However, in general the ranking of techniques according to their period(s) of production will depend on the chosen g and r. Therefore, none of the pragmatic concepts of the period of production gives an unambiguous and generally applicable measure of capital intensity. In the light of the so-called reswitching debate this result is to be expected.

Conclusions

The period of production purports to be a measure of capital intensity. Although it is a useful concept for clarifying the relation between capital and time, it is not, and cannot be, a rigorous measure of aggregate capital per head because even in a restricted model with only one primary input and one consumption output such a measure does not exist. As a *fundamentalist* concept the period of production fails because it cannot simultaneously be a technological concept and explain capital as a subsistence fund; as a *pragmatic* concept it fails because it is not possible to rank techniques according to their period of production independently of the rate of growth and the rate of interest. Hence the period of production cannot avoid the inconsistencies (pointed out in the capital controversies of the 1960s) which are associated with the concept of aggregate capital.

See Also

► Austrian Economics: Recent Work
► Böhm-Bawerk, Eugen Von (1851–1914)

Bibliography

Dorfman, R. 1959. Waiting and the period of production. *Quarterly Journal of Economics* 73: 351–372.

Hicks, J.R. 1939. *Value and capital: An inquiry into some fundamental principles of economic theory.* 2nd ed. Oxford: Clarendon Press.

Kaldor, N. 1937. Annual survey of economic theory: The recent controversy on the theory of capital. *Econometrica* 5: 201–233.

Orosel, G.O. 1979. A reformulation of the Austrian theory of capital and its application to the debate on reswitching and related paradoxes. *Zeitschrift für Nationalökonomie* 29 (1–2): 1–31.

von Böhm-Bawerk, E. 1889. *Kapital und Kapitalzins*. Zweite Abteilung: *Positive Theorie des Kapitals*. Innsbruck: Wagnersche Universitäts-buchhandlung. 4th edn, Jena: Gustav Fischer, 1921. Trans. as *Capital and Interest*, vol. 2: *Positive Theory of Capital,* South Holland, IL: Libertarian Press, 1959.

von Weizsäcker, C.C. 1971. *Steady state capital theory.* Lecture Notes in Operations Research and Mathematical Systems 54. Berlin: Springer.

Weston, J.F. 1951. Some perspectives on capital theory. *American Economic Review: Papers and Proceedings* 41: 129–144.

Periphery

Immanuel Wallerstein

The term ‘periphery’ makes sense only as part of the paired antinomy ‘core(centre)–periphery’. It refers to an economic relationship that has spatial implications. This pair of terms has long been used in the social sciences, but until recently it has been used metaphorically rather than spatially, and to refer to social and political rather than to economic phenomena. Palgrave’s original *Dictionary of Political Economy* (1894–9) did not know the concept.

Nor is it merely an issue of semantics. It is not the case that some other reasonably similar concept had previously been used instead. The issue is more fundamental. Mainstream nineteenth-century economic thought – both classical and neoclassical economics, but to a very large extent Marxism as well – had no place in its theorizing for space, except as location that might affect the cost of a factor of production. Transport costs obviously affected total costs. And location might give a natural rent advantage. Geological deposits were where they were. Water sources that could be dammed for power were located in one place but not another. Space thereupon became one more theoretically accidental, exogenous variable which had to be taken into account in concrete economic practice but was in no sense intrinsic to the functioning of the economic system.

The classic formulation of this view is to be found in the theory of comparative costs. England and Portugal each had certain natural advantages, such that it followed that it was rational, to use Ricardo’s example, for Portugal to exchange her wine for English cloth even though she was able to produce cloth more cheaply than England. In this example the Methuen Treaty never entered the discussion.

It is not that no one ever raised the issue as to whether the natural advantages were not the result of political and social decisions which themselves were integral to the processes of economic behaviour. There had long been, for example, a current of theorizing which justified protectionism. Friedrich List stands out as a leading spokesman of this view in the nineteenth century. The protectionists did argue in effect that comparative advantage was socially structured and that therefore state policy could and should endeavour to transform inequalities. But there are two things to note about this current of protectionist thought. Firstly, it was always marginal to the leading centres of academic economics, and to the extent that its views were incorporated, state policy was once again relegated to the status of an exogenous variable. Secondly, the protectionist current did not challenge, indeed on the contrary it reinforced, a basic pillar of mainstream thought, the parallel and theoretically independent trajectories of a series of states (societies, economies), each of which was separately governed by the same economic laws.

In the interwar period, the worldwide depression in agricultural prices which dates from the early 1920s led to a revival of protectionist theorizing, particularly in those parts of the world which combined three features: a predominance of agricultural production; a small industrial sector; a reasonably large scholarly sector. The three areas which best matched this profile were eastern Europe, Latin America, and India and in all three zones such economic writings appeared. They had in fact, however, rather little impact on local policy and even less on world scholarship.

The situation changed in the post-1945 period. Although the general expansion of the world-economy was no doubt conducive to free trade ideology, the political emergence of the Third World led to some questioning of what in the 1970s would come to be known as 'the international economic order'. It is in this context that the concept of 'periphery' took shape, first of all in the work of Raúl Prebisch and his associates in the UN Economic Commission for Latin America (ECLA).

The original Prebisch thesis laid emphasis on the 'structural' factors which underlay what by the 1950s was being called) 'underdevelopment'. Prebisch argued that peripheral countries were basically exporters of raw materials to industrialized core countries. He argued that there was a long-term decline of the terms of trade against raw materials exporters. Prebisch concluded that this relationship had two basic effects. It maintained the peripheral countries in a vicious cycle of lower productivity and a lower rate of savings than the core countries. And it made it impossible for them to retain the benefits of such increases in productivity as they might experience.

The explanation was 'structural', that is, that there were socio-political 'structures' that affected, even shaped the market, and thereby in (large) part determined advantage in the market. The industrialized countries had 'self-sustained' economies whereas the underdeveloped countries did not, since they functioned as peripheries to centres. The world market forces operated to maintain this undesirable 'equilibrium'. The policy implications were clear. Since the 'normal' operations of the market would only continue the same pattern, state action was required to alter it. The basic immediate recommendation was industrialization via import substitution. The long-run implication was, however, more fundamental. Unlike Ricardo's analysis, the Prebisch argument suggested that the pattern of international trade was established importantly, perhaps primarily, by political decisions and therefore could be changed by political will. Or more generally, the determining framework for the 'world market' was more the overarching world political structure than vice versa.

This basic thesis was picked up and developed by a large number of economists and other social scientists, in Latin America to be sure, but in the Caribbean, in India and Africa as well. It also became the basic argument of a group of social scientists located in Europe and North America, although it should be noted that many of these were persons whose areas of research were in what was now being called the Third World. One of the first of this latter group was H.W. Singer, whose principal contribution was published in 1950, the same year as Prebisch's famous report. For this reason, this viewpoint is sometimes called the Prebisch–Singer thesis.

In time, the Prebisch thesis developed in the 1960s into a doctrine which was called *dependista*, because it emphasized the fact that peripheral areas were in a larger system within which they were 'dependent' as contrasted with more autonomous zones. The primary focus of criticism of the *dependistas* was a dominant mainstream model which was coming to be called 'modernization theory' or 'developmentalism'.

Developmentalism centred around the issue of how those *countries* which were 'underdeveloped' might 'develop'. Developmentalism made several assumptions. Some combination of traits of a country – there was much debate about what they were – led to development. All countries could develop in similar ways, were they to ensure the proper combination of traits – in this sense, the doctrine was melioristic. Development was a patterned process. The last assumption was often expressed as a stage theory. The single most influential expression of this last argument was W.W. Rostow's *Stages of Economic Growth* (1960). Developmentalism originated as an economic doctrine, but others soon began to suggest parallel processes of political development and social development. There was much discussion of the linkages among the various) 'aspects' of development and hence much encouragement of so-called interdisciplinary analysis.

By the 1960s developmentalism had become a dominant and self-conscious mode of analysis in world scholarship, particularly in any discussion of the 'Third World' or the) 'underdeveloped' countries. Prebisch had argued against classical free trade ideology. The main thrust of the 'second generation' of theorizers about the periphery – that of the *dependistas* of the 1960s – was directed against these) 'developmentalists' even though many of them had already accepted the legitimacy of some state intervention in the economy. This second generation was still very largely Latin American – F.H. Cardoso, T. Dos Santos, Celso Furtado, Ruy Mauro Marini, O. Sunkel, R. Stavenhagen were major figures – but there were also Lloyd Best (Trinidad), Samir Amin (Egypt) and Walter Rodney (Guyana). All of these scholars attacked in one way or another the theory of modernization and in particular the assumption that Third World countries could 'repeat' European–North American patterns of development by copying in one way or another the policies, past or present, of the presumably 'successful' states.

The contribution of André Gunder Frank to this second-generation theorizing was that he spelled out two arguments which, while present in the work of his colleagues, had not been as clearly underlined, or as widely disseminated. The first argument is to be found in the slogan he coined, 'the development of underdevelopment'. This is the argument that underdevelopment is not undevelopment, a primordial pre-capitalist or pre-modern state of being, but rather the consequence of the historic process of worldwide development through the linked formation of core and periphery. It followed from this perspective that the further extension and deepening of the division of labour on a world scale led not to national development (as the developmentalists argued) but to the further underdevelopment of the periphery. The policy implications of the two perspectives therefore were directly opposed one to the other.

The second argument involved a critique not of modernization theorists but of so-called orthodox Marxists. To understand this critique we have to look at the history of Marxist theory. From about 1875 on there arose a version of Marxist theory which became predominant in the two major world organizational structures, the Second and Third Internationals, and which very largely reflected the theoretical input of the German Social-Democratic Party (*c*1875–1920) and the Bolsheviks, later Communist Party of the Soviet Union (*c*1900–50). Whether this version was or was not faithful to Marx's own theorizing is not under discussion here, and is irrelevant to the issue at hand.

Since both Internationals were oriented to the issue of obtaining state power, the de facto unit of economic analysis became the state, and, in this respect, there was no real difference with neoclassical models of economic development. Furthermore, under Stalin, a very strong stage model of) 'modes of production' was delineated which paralleled structurally the Rostowian model, although the details were quite different.

In the period 1875–1950, the worldwide structure of capitalist development disappeared or became secondary in) 'orthodox' Marxist theorizing except for a brief interval around World War I where momentarily such figures as Otto Bauer, Nikolai Bukharin, Rosa Luxemburg, and in part Lenin discussed these issues. By the 1920s all such discussion ceased, and by the 1950s Communist parties in Latin America (and elsewhere) were deriving very specific policy implications from the state-centred 'orthodox' theorizing. The reasoning went as follows. Feudalism as a stage comes before capitalism which comes before socialism. Latin America was still in the feudal stage. What was on the politico-economic agenda, and implicitly 'progressive', was national capitalist development. Ergo, Communist parties should enter into political alliances with the national bourgeoisie in order to further national development, postponing to a later date 'socialist revolution'.

The *dependistas* saw this analysis as leading to virtually the same policy results as the analysis of the modernization theory developmentalists. Since the late 1960s was also a period of increasing US–USSR political detente, they saw the theoretical) 'convergence' as tied to a world-level political convergence which in turn was facilitated by the hitherto unremarked common underpinnings of analysis.

The *dependista* popularization of the concept 'periphery' was abetted by two theoretical works which claimed to be Marxist in economic theory yet challenged in each case a major strand in 'orthodox' Marxist economic theorizing. The first was Paul Baran's *Political Economy of Growth*, published in 1957, and which directly inspired many *dependista* authors. Baran modified the concept of *surplus* by introducing a distinction between 'actual' and 'potential' economic surplus, suggesting that the consequence of capitalism was not merely a particular allocation of actual surplus but even more importantly the non-creation of a potential surplus. This non-created potential surplus existed throughout the system but one major component was located in the 'backwardness' of underdeveloped countries.

The second challenge was in Arghiri Emmanuel's *Unequal Exchange*, published in 1969. Emmanuel's book launched a direct attack on the Ricardian theory of comparative advantage, noting that its assumption, the immobility of the factors of production, had never been seriously challenged even by Marxists. Asserting that while capital is internationally mobile, labour has not been, Emmanuel argued that wages determine prices, and not vice versa. Given unequal wages (and immobile labour) internationally, international trade involves unequal exchange, since items priced identically and ensuring parity in rate of profit in fact encompass different amounts of labour. This theory thus challenges the idea that surplus is transferred only in the work process, and that space is irrelevant. The fact that frontiers are crossed is crucial to the theoretical explanation of unequal exchange.

Two other, initially separate intellectual debates entered the scene to complicate the issue further. In the late 1950s, Maurice Dobb and Paul Sweezy had a public debate (in which others then joined) about the so-called transition from feudalism to capitalism in western Europe in early modern times. They disagreed about many things: the time of the change, the motor of change, the geographical context of analysis, the very definition of feudalism and capitalism. What the debate accomplished was that it forced a reconsideration of the definition of feudalism, which was important, since many peripheral zones were being characterized as having) 'feudal' characteristics. When in the late 1950s and 1960s a new debate arose on the nature of, indeed the existence of, an) 'Asiatic mode of production', the debate widened. The more the debate widened, the more the distinction between what is internal and what is external (to the nation/state/society) so fundamental to 'orthodox' Marxist thought, but also to neoclassical thought, came under challenge.

There was a second debate, purely political and far outside world academic circles. It was the obscure, seemingly esoteric debate between the Soviet and Chinese state apparatuses over the process of the hypothetical transition from socialism to communism. This too occurred in the 1950s. The issue was whether states would go forward in this hypothetical transition singly or collectively. This too implied a difference concerning the unit of analysis. The Chinese position had far-reaching implications which by the late 1960s were being called) 'Mao-Zedong thought'.

It was in the 1970s that these strands of thinking about the) 'periphery' and related topics came together. The term) '*dependista*' disappeared. Some began to speak of 'world-systems analysis'. The core–periphery relationship was now being defined as the description of the axial division of labour of the capitalist world-economy. Core and periphery were now less linked locations than linked processes which tended to be reflected in geographical concentrations. These processes had as one major consequence the formation of states within the framework of an interstate system. One could think of the interstate system as the political superstructure of the capitalist world-economy. This world-economy was an historical social system, a socially created whole which developed in specific ways over its history. The overall structure was seen as defining the parameters within which the capitalist market processes occurred. As new geographical zones had been incorporated historically into this system, they had been for the most part ‘peripheralized’. This meant that various worldwide mechanisms (political, financial, and cultural) tended to make it profitable for individual entrepreneurs to segregate production processes spatially such that some zones had disproportionately high concentrations of peripheral processes – that is, processes with a high labour component and relatively low-cost labour – ensured by the involvement of wage-workers in these zones in usually reorganized household structures in which lifetime income returns from wage labour comprised a minority percentage of total real revenue.

While state policies could affect these relationships, the ability of any single state to transform the situation was constrained by its location in the interstate system and therefore depended significantly upon the changing condition of the balance of power. The interstate system varied in patterned ways between periods in which there was one hegemonic power and periods in which there was acute rivalry among several strong powers.

In addition, the ability of states to affect the processes of peripheralization was said to be a function of the cyclical rhythms of the world-economy, believed to alternate, once again in patterned ways, between periods of expansion and stagnation.

The regular cyclical rhythms and the alterations of the conditions of the interstate system led to some continuous but limited shifting in the economic roles of particular geographical zones within the system without necessarily changing the basic structuring of core–periphery relations.

Finally, it has been argued that the geographical concentration of different economic processes has been trimodal rather than bimodal, there having been at all times semiperipheral zones, defined as regions having a fairly even mix of core-like and periphery-like economic processes.

The concept ‘periphery’ thus has involved a basic theoretical criticism of nineteenth-century economic paradigms. It has not been spared counterattack from three main quarters: of course from the modernization/developmentalists under attack, most of whom have been basically Keynesians in their economic theorizing; but even more from so-called neo-liberals (the critique of P.T. Bauer has been the most trenchant), and from) ‘orthodox’ Marxists.

The concept ‘periphery’ has served a polemical purpose in the last 20 years. To advance its utility, its proponents must now come to clearer terms about the functioning interrelations of the three antinomies of the capitalist world-economy; core–periphery relations in the division of labour; A and B phases in the cyclical long waves; and periods of hegemony versus periods of rivalry in the interstate system.

See Also

- ▶ North–South Economic Relations
- ▶ Terms of Trade
- ▶ Unequal Exchange
- ▶ Uneven Development

Bibliography

Amin, S. 1974. *Accumulation on a world scale*. New York/London: Monthly Review Press.

Arrighi, G. 1983. *The geometry of imperialism*, Revised edn. London: Verso.

Baran, P. 1957. *The political economy of growth*. New York: Monthly Review Press.

Bauer, P.T. 1972. *Dissent on development*. Cambridge, MA: Harvard University Press.

Emmanuel, A. 1969. *Unequal exchange*. New York/London: Monthly Review Press, 1972.

Frank, A.G. 1969. *Latin America: Underdevelopment or revolution*. New York/London: Monthly Review Press.

Furtado, C. 1963. *The economic growth of Brazil*. Berkeley/Los Angeles: University of California Press.

Hilton, R., ed. 1976. *The transition from feudalism to capitalism*, Revised edn. London: New Left Books.

Hirschman, A.O. 1958. *The strategy of economic development*. New Haven/London: Yale University Press.

Hopkins, T.K., and I. Wallerstein. 1982. *World-systems analysis*. Beverly Hills: Sage.

Love, J.L. 1980. Raúl Prebisch and the origins of the doctrine of unequal exchange. *Latin American Research Review* 15(1): 45–72.

Prebisch, R. 1950. *The economic development of Latin America and its principal problems*. New York: United Nations.

Rostow, W.W. 1960. *The stages of economic growth*. New York/London: Cambridge University Press.

Singer, H.W. 1950. The distribution of gains between investing and borrowing countries. *American Economic Review* 40(2): 473–485.

Wallerstein, I. 1974, 1980. *The modern world system*, 2 vols. New York/San Francisco/London: Academic Press.

Perlman, Selig (1888–1959)

M. Donnelly

Keywords
Commons, J. R.; Perlman, S.; Socialism; Trade unionism

JEL Classifications
B31

Selig Perlman was born in 1888 at Bialystok, Poland, then a part of Tsarist Russia. His father was a yarn spinner. Perlman grew up in an atmosphere shaped at once by the labour movement, socialism and Zionism. He emigrated to the United States in 1908, and took up studies with John R. Commons at the University of Wisconsin. He received his Ph.D. there in 1915, joined the teaching faculty and became professor of economics in 1927. He collaborated on the four-volume *History of Labor in the United States* compiled by Commons (1918–35), and published *A History of Trade Unionism in the United States* (1922). His most important work, the influential *A Theory of the Labor Movement,* appeared in 1928. Perlman died in 1959.

Perlman's early sympathies were Marxist, but his views were shaken considerably upon his going to America and coming under the influence of Commons. He came to regard the ideas of socialism as essentially the creation of intellectuals, fundamentally at odds with manual workers' own aspirations and experience. Where the labour movement is weak, Perlman argued, it is more susceptible to control by intellectuals; where, on the other hand, political conditions allow it to become strong, the labour movement is better able to outgrow its early ideological trappings and advance to maturity. Late 19th- and early 20th-century America seemed to Perlman the clearest case of a labour movement 'emancipated from the hegemony of intellectual revolutionists' and expressing its own 'philosophy of organic labor'. The key to successful trade unionism, in Perlman's view, was a limited, practical 'job-consciousness', struggling toward collective control of employment opportunities but not otherwise challenging the prerogatives of capitalists.

Selected Works

1918–35. (With J.R. Commons et al.) *History of labor in the United States.* New York: Macmillan.

1922. *A history of trade unionism in the United States.* New York: Macmillan.

1928. *A theory of the labor movement*. New York: Macmillan.

Permanent-Income Hypothesis

Mark Aguiar and Erik Hurst

Abstract
The permanent income hypothesis (PIH) is a theory that links an individual's consumption at any point in time to that individual's total income earned over his or her lifetime. The hypothesis is based on two simple premises: (1) that individuals wish to equate their expected marginal utility of consumption across time and (2) that individuals are able to respond to income changes by saving and dis-saving. In this article we present the intuition and empirical implications of the PIH in several standard contexts.

Keywords
Buffer stocks; Consumption insurance; Euler equations; Impatience; Liquidity constraints; Marginal utility of consumption; Martingales; Permanent income hypothesis; Precautionary

wealth; Preferences; Retirement; Retirement consumption puzzle; Uncertainty

JEL Classifications
D4; D10

The permanent income hypothesis (PIH) is a theory that links an individual's consumption at any point in time to that individual's total income earned over their lifetime.

The PIH is based on two simple premises: (1) that individuals wish to equate their expected marginal utility of consumption across time and (2) that individuals are able to respond to income changes by saving and dis-saving. Because consumers are making their consumption decisions based on lifetime resources, the PIH implies that today's consumption will respond differently to changes in today's income depending on whether the income changes are expected as opposed to unexpected, or temporary as opposed to permanent. The PIH provides a sharp contrast to Keynesian consumption rules, which assume consumers make their consumption decisions based only upon current income.

The major insights of the PIH originated in Friedman (1957). They are closely related to the ideas expressed in Modigliani and Brumberg's (1954) life-cycle hypothesis (see Carroll 2001, for a summary of Friedman's original work). Since the 1950s there have been many additional theoretical and empirical contributions. This article presents the intuition and empirical implications of the PIH that have evolved since the 1950s in several standard contexts.

The Canonical Model

Consider the canonical model in which an individual lives $T+1$ periods and earns y_t in period $t = 0,\ldots,T$. For now, we assume that the income stream is known at time zero. The canonical model assumes that the individual can borrow and lend freely at an interest rate r. The standard model also assumes that the future is discounted at the rate $\beta < 1$ and utility is additively separable across time and additively separable across consumption and leisure. For simplicity, we treat leisure as fixed and treat income as exogenous to the consumer. We revisit these assumptions below. Let $u(c)$ represent the period utility enjoyed from consumption, where $u' > 0$; $u'' < 0$. The consumer's problem is therefore:

$$\max_{\{c_t\}_{t=0}^{T}} \sum_{t=0}^{T} \beta^t u(c_t) \tag{1}$$

subject to $\sum_{t=0}^{T}(1+r)^{-t}c_t \leq \sum_{t=0}^{T}(1+r)^{-t}y_t + A_0$, where A_0 represents initial assets.

A necessary condition for an interior optimal consumption plan is $u'(c_t) = \beta(1+r)u'(c_{t+1})$, for all $0 \leq t \leq T-1$. Therefore, the relationship between consumption in two periods is independent of the relationship between income in those two periods. For example, suppose that individual's discount the future at the rate of interest such that $\beta(1+r) = 1$. With such a restriction on preferences, the individual will consume the same amount each period. Also for simplicity, let $T \to \infty$ and $A_0 = 0$ (and impose the 'no-Ponzi-game' condition $lim_{t\to\infty}A_t/(1+r)^t \geq 0$). The budget constraint then implies that consumption in each period equals the annuity value of the present discounted value of income, or 'permanent income,' such that:

$$c = r\sum_{t=0}^{\infty}(1+r)^{-t}y_t. \tag{2}$$

Note that consumption is a function only of permanent income, and not how that income is allocated across periods. The ability to borrow and lend is key to the permanent income hypothesis. This allows the individual to transfer income across periods at the rate $(1+r)$. Access to such an asset makes the present discounted value of income the only relevant constraint on consumption.

The result has a natural implication in a life-cycle model. Suppose individuals work for $S < T$ periods and then retire. Aside from a potential trend due to time discounting, the PIH implies that consumption should not respond to the drop in income at a known period of retirement. Rather, assets built up over the working years are used to finance retirement consumption. Similar examples are plentiful. For example, a teacher on a 9-month salary consumes steadily over 12 months, or a yearend bonus is used for purchases throughout the year. The fact that income is expected to change tomorrow should already be incorporated into today's consumption plan.

In the above model, there was no uncertainty about future income. This is reasonable for predictable changes to income such as retirement or seasonal work, but less useful in understanding consumption's response to unexpected 'shocks' such as an unemployment spell or changes in business cycle conditions. We extend the model to the case of uncertainty by assuming that income follows a stochastic process. In particular, let *yt* denote the random variable of income at time $t = 0,\ldots,T$.

We continue our assumption that individual's have access to a risk-free bond. Let *Et* denote expectations conditional on information as of time *t*. At any point in time, *t*, the consumer's problem can be expressed as the following:

$$\max_{\{c_\tau\}_{\tau=t}^{T}} E_t \sum_{\tau=t}^{T} \beta^{\tau-t} u(c_\tau) \qquad (3)$$

subject to the period-by-period budget constraint: $A_{t+1} = (1 + r)(A_t + y_t - c_t)$. Notice that Eq. (3) differs from Eq. (1) in that individuals in Eq. (3) are maximizing expected utility. The first-order conditions imply the following 'Euler equation':

$$u'(c_t) = \beta(1 + r)E_t u'(c_{t+1}). \qquad (4)$$

The marginal utility of consumption varies in a predictable way due only to the interest rate and the subjective discount rate. All other movements are unpredictable (with respect to information available prior to time *t*). Jensen's inequality implies that consumption will be a martingale when $\beta = 1 + r$ only if marginal utility is linear in consumption (that is, quadratic utility). In many standard utility functions, marginal utility is convex, implying that consumption trends upward in expectation when marginal utility is a martingale. Moreover, all else equal, consumption will respond more to unanticipated permanent innovations to income than to transitory innovations.

Empirical Tests of the Canonical Model

Equation (4) states that, aside from r and β, information known at time *t* should not affect the change in the marginal utility of consumption between *t* and $t + 1$. Estimating Eq. (4) has been the focus of numerous empirical studies, beginning with seminal paper of Hall (1978). Using aggregate data, Hall finds that lagged consumption and lagged income have minimal predictive power for changes in current consumption growth between *t* and $t + 1$. This, by itself, may be interpreted as a victory for the PIH. However, Hall also finds that a lagged index of stock prices does have predictive power for future consumption changes, an apparent violation of Eq. (4). Hall's study was followed by a large empirical literature exploiting aggregate consumption data to test whether innovations to consumption are predictable using information available in prior periods. However, a consensus has emerged that aggregation issues undermine the validity of tests using aggregate data.

A large literature has emerged testing Eq. (4) using micro data. For example, Attanasio and Weber (1995) and Attanasio and Browning (1995) find support for the PIH using data from the US Consumer Expenditure Survey and the UK Family Expenditure Survey, respectively. Additionally, Shea (1995), Parker (1999), Souleles (1999), Browning and Collado (2001), and Hsieh (2003), among others, have used micro data to examine how consumption responds to anticipated changes in income. These results, however, have been mixed. The conclusion of this literature is that, at least in some instances, consumption responds to predictable changes in income. This excess sensitivity of consumption to predictable income changes has been seen as a violation of the canonical model of the PIH outlined above.

Moving Beyond the Canonical Model

Depending on the context, the ability to freely borrow and lend may be considered too restrictive or not restrictive enough. On the one hand, it rules out state-contingent insurance contracts between consumers. On the other hand, the ability to borrow against future income is often limited in practice due to lack of enforcement. We now briefly describe how the canonical PIH differs from optimal consumption patterns in models with complete insurance markets or models with borrowing constraints.

Perfect insurance in an economy inhabited by agents that enjoy utility as given by Eq. (3) implies that individual consumption depends only on aggregate income rather than how that income is distributed across individuals. That is, consumption depends only on aggregate shocks and not on idiosyncratic shocks. This contrasts with the PIH's statement that consumption responds to idiosyncratic permanent income shocks. The difference reflects the limits of the insurance provided by a risk-free bond. However, there is a parallel as noted by Cochrane (1991). The implication that consumption should not respond to idiosyncratic income shocks was formalized and tested by Townsend (1994) using data from Indian villages and Cochrane (1991) using US data. While Townsend rejects perfect risk sharing, he presents evidence that there is significant insurance of idiosyncratic shocks within villages in India. Cochrane rejects perfect insurance in the case of long illness and involuntary job loss, but fails to reject in the case of several other idiosyncratic shocks.

Another alternative to the standard PIH asset market structure is limiting the amount one can borrow against future income. The inability to borrow implies that Eq. (4) may not hold. When constrained, a consumer may be forced to adjust consumption in response to a transitory or predictable shock to income. For example, if an individual receives a temporary income decline, the inability to borrow against future income may necessitate that consumption moves with contemporaneous income. Zeldes (1989) argues that liquidity constraints do bind for a significant fraction of consumers. Moreover, the inability to borrow presents consumers with the risk that a series of negative income shocks may force consumption down to extremely low levels. To mitigate this risk, potentially constrained consumers build up a 'buffer stock' of savings. See precautionary saving and precautionary wealth for a discussion of the accumulation of wealth for precautionary reasons.

Life-Cycle Consumption

While liquidity constraints can explain the empirical fact that consumption is excessively sensitive to changes in predictable income, empirical critiques remain about the ability of individuals to rationally make consumption decisions today based on their expectations of future income realizations. Two of the strongest critiques are that consumption expenditures are hump-shaped over the life cycle (peaking when households are in their mid-forties) and that there is a significant decline in consumption expenditures at the time of retirement. The latter fact has been referred to as the 'retirement consumption puzzle' and has been documented and discussed by, among others, Bernheim, Skinner and Weinberg (2001).

The two empirical critiques are related. According to the standard permanent income hypothesis outlined above, individuals should be smoothing their marginal utility of consumption over their lifetimes. Researchers have been trying to modify the PIH so that it matches these two additional empirical facts. For example, Attanasio et al. (1999) find that, if preferences are a function of demographics, the life-cycle profile can be matched. Alternatively, Gourinchas and Parker (2002) find that a model with a properly calibrated income process can match the hump-shaped consumption profile if households are liquidity constrained and sufficiently impatient.

Aguiar and Hurst (2005, 2007) adopt a different approach from those above by appealing to the intuition of Becker (1965). They argue that the PIH theory concerns consumption while the data reports expenditure. The distinction is important because consumption requires time as well as market goods. In particular, households may substitute time for expenditure and maintain a constant level of consumption as expenditures fall. This margin of substitution is suppressed in the canonical form of the model, but Aguiar and Hurst (2005, 2007) document that it is empirically important and reconciles the PIH with both the life-cycle profile of expenditure and the changes in expenditure associated with retirement.

In summary, the current state of literature has expanded on the insights of Friedman's original discussion of the PIH by building in additional features to the canonical model to match a wide variety of empirical regularities. However, this discussion highlights the broader point that any empirical test of the PIH is always a joint test of the hypothesis itself as well as the specific restrictions the researcher places on preferences (for example, whether utility is non-separable between consumption and leisure, the curvature of marginal utility, or the extent to which individuals are impatient), information (for example, assumptions about the income process), or technologies (for example, the existence of liquidity constraints, a home production sector, or complete markets) used to construct the hypothesis' empirical counterpart.

See Also

- ▶ Friedman, Milton (1912–2006)
- ▶ Modigliani, Franco (1918–2003)
- ▶ Precautionary Saving and Precautionary Wealth

Bibliography

Aguiar, M., and E. Hurst. 2005. Consumption vs. expenditure. *Journal of Political Economy* 113: 919–948.

Aguiar, M., and E. Hurst. 2007. Lifecycle prices and production. *American Economic Review* 97(5): 1533–1559.

Attanasio, O., and M. Browning. 1995. Consumption over the life cycle and over the business cycle. *American Economic Review* 85: 1118–1137.

Attanasio, O.P., and G. Weber. 1995. Is consumption growth consistent with intertemporal optimization? Evidence from the consumer expenditure survey. *Journal of Political Economy* 103: 1121–1157.

Attanasio, O., J. Banks, C. Meghir, and G. Weber. 1999. Humps and bumps in lifetime consumption. *Journal of Business and Economic Statistics* 17: 22–35.

Becker, G. 1965. A theory of the allocation of time. *Economic Journal* 75: 493–508.

Bernheim, B.D., J. Skinner, and S. Weinberg. 2001. What accounts for the variation in retirement wealth among U.S. households? *American Economic Review* 91: 832–857.

Browning, M., and D. Collado. 2001. The response of expenditures to anticipated income changes: Panel data estimates. *American Economic Review* 91: 681–692.

Carroll, C. 2001. A theory of the consumption function, with and without liquidity constraints. *Journal of Economic Perspectives* 15: 23–45.

Cochrane, J. 1991. A simple test of consumption insurance. *Journal of Political Economy* 99: 957–976.

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

Gourinchas, P.-O., and J. Parker. 2002. Consumption over the life-cycle. *Econometrica* 70: 47–89.

Hall, R. 1978. Stochastic implications of the life cycle-permanent income hypothesis: theory and evidence. *Journal of Political Economy* 86: 971–987.

Hsieh, C.-T. 2003. Do consumers react to anticipated income changes? Evidence from the Alaska permanent fund. *American Economic Review* 93: 397–405.

Modigliani, F., and R. Brumberg. 1954. Utility analysis and the consumption function: An interpretation of the cross section data. In *Post-Keynesian economics*, ed. K. Kurihara. New Brunswick: Rutgers University Press.

Parker, J. 1999. The reaction of household consumption to predictable changes in payroll tax rates. *American Economic Review* 89: 959–973.

Shea, J. 1995. Union contracts and the life-cycle/permanent income hypothesis. *American Economic Review* 85: 186–200.

Souleles, N. 1999. The response of household consumption to income tax refunds. *American Economic Review* 89: 947–958.

Townsend, R. 1994. Risk and insurance in village India. *Econometrica* 62: 539–591.

Zeldes, S. 1989. Consumption and liquidity constraints: An empirical investigation. *Journal of Political Economy* 97: 305–346.

Perron–Frobenius Theorem

Hukukane Nikaido

Keywords

Brouwer's fixed point theorem; Hawkins-Simon conditions; Leontief system; Perron-Frobenius theorem

JEL Classifications

C0

A linear transformation mapping $(x_1,x_2, \ldots, x_n)$ to $(y_1, y_2, \ldots, y_n)$ by

$$y_i = \sum_{j=1}^{n} a_{ij}x_j(i = 1, 2, \ldots, n)$$

all of whose coefficients a_{ij} are non-negative has special properties not shared by the general linear transformation. In a matrix form the transformation takes the form

$$x \to y = Ax,$$

where A is the n-dimensional square matrix with elements a_{ij} in the ith row and the jth column, x is the vector having x_j in the jth component and y is the vector y_i in the ith component. The non-negativity of elements of the coefficients matrix A obviously implies that a vector x with all its components non-negative is mapped to a vector y with all its components non-negative in the transformation. This peculiar nature gives rise to special properties of the eigenvalues and associated eigenvectors of the matrix A. Among them, those found and proved by Frobenius (1908, 1909, 1912), also already noticed for a special case by Perron (1907), are the most relevant to linear economic models in which variables are non-negative. The Perron–Frobenius theorem states them in several propositions

(1) A has real non-negatives eigenvalues. With the largest $\lambda = \lambda(A)$ of the non-negative eigenvalues is associated an eigenvalues x having non-negative components fulfilling

$$\lambda x = Ax.$$

(2) The absolute value $|\omega|$ of any eigen value ω of A, either real or complex, is bounded by $\lambda(A)$ so that $|\omega| \leq \lambda(A)$.

(3) The matrix $\rho I - A$ where I is the identity matrix and ρ is a real number, has an inverse matrix with all its elements non-negative if and only if ρ is larger than $\lambda(A)$.

Alternative methods of proving the propositions (1), (2) and (3) are available. Some of them are given below.

Proof of (1). The proof is straightforward for A with all its elements positive. Among all the pairs (θ, y) of a real number θ and a nonzero vector y having all its components non-negative that fulfil the n inequalities, the ith component of A$y \geq$ the ith component of $6y$ ($i = 1, 2, n$) there is one (λ, x) with λ being the largest of all such θ. Then $Ax = \lambda x$. For otherwise, some components are larger than the corresponding components of λx while the other components of Ax are not less than the corresponding ones of λx. Whence all the components of $A(Ax)$ are larger than the corresponding ones of λ (Ax) so that λ can be further increased to get another pair (θ, Ax), $\theta > \lambda$ fulfilling the inequalities, contrary to the maximum property of λ. Generally A can be approximated from above by $A_\varepsilon = A + \varepsilon T$ where ε is a small positive number and T is the matrix all of whose elements are one. A_ε with all its elements positive has a special pair $(\lambda_\varepsilon, x_\varepsilon)$ satisfying $A_\varepsilon x_\varepsilon = \lambda_\varepsilon x_\varepsilon$ and maximizing θ at λ_ε over all pairs (θ, y) fulfilling the inequalities, the ith component of $A_\varepsilon y \geq$ the ith component of θy ($i = 1, 2, ...,$ n). Then, for $\varepsilon' \geq \varepsilon$ the ith component of $A_{\varepsilon'} x_\varepsilon \geq$ the ith component of

$$A_\varepsilon x_\varepsilon = \lambda_\varepsilon x_\varepsilon (i = 1, 2, \ \ldots, \ n),$$

implying $\lambda_{\varepsilon'} \geq \lambda_\varepsilon$, Whence, λ_ε converges monotonically to a non-negative λ as ε decreases toward zero. The corresponding eigenvector x_ε if so normalized that its components sum up to 1, converges to a nonzero vector having all its components non-negative for a subsequence $\varepsilon(s)$ of positive numbers tending monotonically to zero when s $\to \infty$. Hence $A_{\varepsilon\ (s)} x_{\varepsilon(s)} = \lambda_{\varepsilon(s)} x_{\varepsilon(s)}$ becomes $Ax = \lambda x$ in the limit when $s \to \infty$.λ is the largest of θ of all pairs (θ, y) fulfilling the inequalities, the ith component of $Ay \geq$ the ith component of θy ($i =$ 1, 2, ..., n). For $\lambda_\varepsilon \geq \theta$ by construction, which becomes $\lambda \geq \theta$ in the limit.

Alternatively, this proposition can be proved by virtue of Brouwer's fixed point theorem as a fixed point of the mapping that transforms each vector x with non-negative components $x_i (i = 1, 2, \ \ldots, n)$ adding up to unity to a vector y with components

$$y_i = \left(x_i + \sum_{j=1}^{n} a_{ij} x_j \right) \Big/ \left(1 + \sum_{k,j=1}^{n} a_{kj}, x_j \right),$$
$$(i = 1, 2 \ldots . n)$$

At a fixed point x^* that is transformed to itself these equations can be rearranged to

$$\lambda x^* = Ax^*,\ \lambda = \sum_{k,j=1}^{n} a_{kj}x_j^*.$$

Then λ (A) is obtained as the largest of λ's of all such fixed points.

Proof of (2). For an eigenvalue ω of A and an associated eigenvector z with components z_j, the equations

$$\omega z_i = \sum_{j=1}^{n} a_{ij}, z_j, (i = 1, 2, ..., \quad n)$$

hold by definition. Then the absolute values of ω, z_i satisfy

$$\sum_{j=1}^{n} a_{ij}|z_j| \geq |\omega||z_i|, \quad (i = 1, 2, \quad ..., \quad n)$$

Whence λ (A) ≥ |ω| by the maximum property of λ (A). In particular, λ (A) is the largest of all non-negative eigenvalues of A.

Proof of (3). Necessity. If $\rho I - A$ has an inverse matrix having all its elements non-negative, $p'(\rho I - A)$ has all its elements positive for some vector ρ having all its components positive, where the prime stands for transposition. Then, $Ax = \lambda x$, $\lambda = \lambda(A)$ with x an associated eigenvector having all its components non-negative, becomes, when pre-multiplied by p' , $\rho p'x$, $\lambda p'$ x , p' $x > 0$, which implies $\rho > \lambda$.

Sufficiency. First note that $\lambda(A) \geq \lambda(C)$ for any principal minor matrix C of A. For, if $\lambda(C)\,y = Cy$ for an eigenvector y with all its components non-negative associated with λ(C) the inequalities, the ith component of $Az \geq$ the ith component of $\lambda(C)$ $z(i = 1, ,2, ..., n)$ hold for the vector z augmented from y by putting zero in the missing components, so that λ (A) ≥ $\lambda(C)$ by the maximum property of λ (A). *If* $\rho > \lambda = \lambda(A)$, the determinant of $\rho I - A$ must not be zero, for otherwise ρ would be a positive eigenvalue of A larger than λ (A). Hence ρI–A is nonsingular and invertible. For any vector c with non-negative components $x_i = (\rho I - A)^{-1}c$ must have all its components non-negative. Otherwise x would have some components negative, and an identical, simultaneous renumbering of equations and variables would bring the relation between x and c to the form

$$\begin{aligned} \rho x_i - \sum_{j=k+1}^{n} a_{ij}x_j &= \sum_{j=1}^{k} a_{ij}x_j + C_j, & (i = k+1, \ldots, n) \\ x_j &\geqq 0, & (j = 1, 2, \ldots, k) \\ x_j &< 0, & (j = k+1, \ldots, n) \end{aligned}$$

which are non-negative on the right side. Whence

$$\begin{aligned} &\sum_{j=k+1}^{n} a_{ij}y_i \geqq \rho y_i, (i = k+1, \ldots, n) \\ &y_i = -x_j > 0, \quad (j = k+1, \ldots, n) \end{aligned}$$

so that $\rho > \lambda(A) \geq \lambda(C)$ contrary to the maximum property of λ(C) for the principal minor matrix C of A obtained by deleting the first k rows and columns of A. This shows that the components of x are non-negative, which ensures the non-negativity of all the elements of $(\rho I - A)^{-1}$.

The condition in (3) that ρI –A has an inverse matrix with all its elements non-negative can be paraphrased as the positivity of all the principal minor determinants of ρI –A the so-called *Hawkins–Simon conditions.*

The Perron–Frobenius theorem pertains to the possibility of special solutions of linear economic models and to the 'good behaviour' of those solutions. The most typical instance of such models is the Leontief system. In a Leontief system consisting of n sectors, each of which produces a single good, without joint products, under constant returns to scale, using n goods as current input and as capital, let a_{ij} and b_{ij} be the amounts of the ith good consumed as input and used as capital, respectively, which are necessary to produce one unit of the jth good in the jth sector (i, j = 1,2, ..., n).

Let the levels of sectoral output $x_j(t)$ at time t (j =1,2, ..., n) be so determined that net outputs are invested to increase capital. Then

$$x(t) = Ax(t) + B(x(t+1) - x(t)),$$

where A and B are the input coefficients matrix and the capital coefficients matrix having elements a_{ij}

and b_{ij} in the *i*th row and the *j*th column, respectively. A special time path of output $x_i(t) = (1 + g)^t x_i$ ($i = 1,2, ..., n$), called a balanced growth path, on which the levels of sectoral output grow at the equal positive rate g is generated by an eigenvector x with non-negative components x_i associated with the Perron–Frobenius eigenvalue λ (=1/g) of the matrix $(I - A)^{-1}B$ having all elements non-negative, provided the system is *productive* enough for A to have its Perron–Frobenius eigenvalue less than 1. On the dual side a row eigenvector p' with non-negative components $p_j (j = 1, 2, \ldots, n)$ associated with the Perron–Frobenius eigenvalue $\lambda(= 1/r)$ of $B(I - A)^{-1}$, equal to that of $(I - A)^{-1}B$, gives a special set of prices determined by

$$p' = p'A + rp'B$$

at which the sectoral rates of profit are equalized to the common rate $r = 1/\lambda$.

In the system of Sraffa (1960) in which input–output correspondences are

$$\begin{aligned}&(A_a,\ B_a,\ \ldots, K_a) \to A(A_b,\ B_b,\ldots, K_b)\\&\times \to B(A_k,\ B_k,\ldots, K_k) \to K\end{aligned}$$

the standard commodity is constructed by non-negative multipliers $q_a, q_b,\ldots, q_k$ that fulfil

$$\begin{aligned}&(A_a q_a + A_b q_b\ \ldots\ + A_k q_k)(1 + R)\\&= A q_a (B_a q_a + B_b q_b +\ \ \ldots\ \ + B_k q_k)(1 + R)\\&= B q_b (K_a q_a + K_b q_b +\ \ \ldots + K_k q_k)(1 + R) = K q_k.\end{aligned}$$

These multipliers are obtained as components of an eigenvector associated with the Perron–Frobenius eigenvalue λ[=1/(1 + R)] of the matrix

$$\begin{pmatrix} A_a/A & A_b/A & \ldots A_k/A \\ B_a/B & B_b/B & \ldots B_k/B \\ K_a/K & K_b/K & \ldots K_k/K \end{pmatrix}.$$

More specific information is available about the Perron–Frobenius eigenvalue of those matrices having all their elements non-negative in such a way as to be *indecomposable,* in the sense that no identical, simultaneous renumbering of its rows and columns can be put into the form

$$\begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix},$$

where A_{11} and A_{22} are square submatrices while A_{12} is a rectangular submatrix and 0 is a rectangular submatrix having zero in all its elements.

(4) If A is indecomposable, the Perron–Frobenius eigenvalue $\lambda(A)$ is positive, and with it is associated an eigenvector x having all its components positive. Any eigenvector associated with $\lambda(A)$ is a scalar multiple of x.

(5) $\lambda(A)$ is a simple root of the characteristic equation. If A has s eigenvalues of moduli equal to $\lambda(A)$, they give all the roots of the equation

$$\omega^s = \lambda(A).$$

(6) By an identical, simultaneous renumbering of the rows and columns A can be put in a form

$$\begin{bmatrix} 0 & . & . & . & 0 & A_{15} \\ A_{21} & 0 & & & & 0 \\ 0 & A_{32} & 0 & & & \\ . & & & & & \\ . & & & & & \\ 0 & . & . & 0 & A_{35-1} & 0 \end{bmatrix},$$

where s is the number of eigenvalues of moduli equal to $\lambda(A)$ and $A_{1s}, A_{21}, \ldots, A_{ss-1}$ are rectangular submatrices, while all the other elements are zero.

A standard reference compiling the main results centring around the Perron–Frobenius theorem is Debreu and Herstein (1953).

See Also

▶ Linear Models

Bibliography

Debreu, G., and I.N. Herstein. 1953. Nonnegative square matrices. *Econometrica* 21: 597–607.

Frobenius, G. 1908. *Über Matrizen aus positiven Elementen I. Sitzungsberichte der Königlich-Preussischen Akademie der Wissenschaften.*

Frobenius, G. 1909. *Über Matrizen aus positiven Elementen II. Sitzungsberichte der Königlich-Preussischen Akademie der Wissenschaften.*

Frobenius, G. 1912. *Über Matrizen aus nichtnegativen Elementen. Sitzungsberichte der Königlich-Preussischen Akademie der Wissenschaften.*

Perron, O. 1907. Zur Theorie der Matrizen. *Mathematische Annalen* 64: 248–263.

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

Perroux, François (1903–1987)

Henry W. Spiegel

Keywords

Domination effect; Economic space; Inequality; Market power; Perroux, F.; Poles of development; Regional development; Structural change; Unbalanced economic growth

JEL Classifications

B31

French economist, best known for his construction of a theoretical system of economic power. He was born in Lyon and his academic career led him to the Sorbonne and from 1955 to 1975 to the Collège de France. A critic of neoclassical economics, Perroux shared some of the concerns of the American institutionalists, but went beyond them by constructing a system of economic analysis – the only one at his time – that rivals conventional equilibrium economics. This system, comprehensive and consistent, is grounded in an all-pervasive 'domination effect' that reflects the inequality of economic agents with respect to their economic power. In equilibrium economics the actions of the economic agents are considered coordinated by an adequate amount of equality, leading to mutual concessions that in turn bring about adjustments and the removal of disturbances. Such an approach, according to Perroux, is contradicted by the facts of economic life and fails to reveal the role of economic power in the market. Where conventional economics stresses coordination among equals and their functional interdependence, Perroux sees a relationship of subordination among economic agents, with the latter either dominating or dominated. Just as Schumpeter, who influenced Perroux and about whom he wrote a book, had revealed the dynamics of innovation, so Perroux disclosed the dynamics of inequality. He acknowledged that there were other theories of monopolistic market situations that shared features of his own ideas, and himself introduced Chamberlin's work to French readers, but pointed out that these theories covered only special cases that would be more adequately handled by a general theory such as that developed by him.

Perroux described his domination effect as asymmetrical and irreversible and as not presuming any intention on the part of the dominating agent. Unlike conventional economics, the domination effect does not produce equilibrium but protracted and cumulative changes. It operates at the level of the firm, of the industry and of the national economy. A dominant firm, for example, can integrate its operations and earn a surplus from increasing sales to and declining purchases from the outside, and from a market position that yields it favourable prices. The surplus adds to the power of the dominant firm by providing it with means for internal financing, for mergers and acquisitions, and for financing or manipulating the demand for its products. At the international level Perroux's domination effect yields new insight into the position of the dominant economy. His theory differs from the theories of imperialism by not requiring an intention on the part of the dominating power.

Perroux's general theory of economic power was developed during the 1940s and 1950s, not long after the new theories of Keynes, input–output analysis, mathematical programming and game theory had been absorbed into mainstream economics. The economics profession was not ready for still another profound change. Thus, Perroux's general theory of economic domination did not upset conventional analysis. However, from his general theory Perroux derived theories of 'economic space' and 'poles of development', which in turn yielded theories of structural change, unbalanced economic growth and regional development that continue to be widely discussed and applied in regional planning.

Selected Works

1950a. Economic space: Theory and applications. *Quarterly Journal of Economics* 64(1): 89–104.

1950b. The domination effect and modern economic theory. *Social Research* 17(2): 188–206. 1969. *L'économie du XXe siècle,* 3rd edn. Paris: PUF.

Bibliography

Bocage, D. 1985. *The general economic theory of François Perroux*. Lanham: University Press of America.

Perroux, F. 1980. Peregrinations of an economist and the choice of his route. *Banca Nazionale del Lavoro Quarterly Review* 33 (June): 147–162.

Uri, P. 1984. Perroux. In *Contemporary economists in perspective*, ed. H.W. Spiegel and W.J. Samuels, vol. 2. Greenwich: JAI Press.

Personal Debt and Psychological Health

John Gathergood

Abstract

Problematic personal debts and associated outcomes, such as bankruptcy and foreclosure, lead to significant declines in psychological health. This article summarises the recent literature and discusses the key issues in measurement and causality. Medical studies show that problem debts are associated with depression, self-harm and even suicide. Recent studies using econometric techniques show that some of the association in self-reported data is due to perception bias. Quasi-experimental studies using data from the housing crises show the onset of problem debt causes deterioration in psychological health, including effects upon physical health and health behaviours.

Keywords

Anxiety; Bankruptcy; Consumer credit; Debt; Depression; Finance; Foreclosure; Health; Health behaviours; Mental illness; Mortgage debt; Psychological health; Repossession; Suicide

JEL Classification

I310; D1; D140

Introduction

Problematic personal debts and associated outcomes, such as bankruptcy and foreclosure, lead to significant declines in psychological health. High levels of consumer debt are associated with anxiety, depression and poor general health. Personal bankruptcy or foreclosure events are associated with large declines in psychological health and increased likelihood of adverse health outcomes, including stress-related medical conditions, self-harm and suicide.

Economists are interested in the potential negative effects of personal debt because in standard economic models access to debt is welfare-improving. In dynamic consumption theory access to debt allows consumers to smooth their (marginal utility of) consumption over time and hence increase overall utility. However, inability to repay debt can result in consumer bankruptcy or foreclosure on property, which incurs large welfare losses to individual consumers through loss of goods and services, and the experience of these may impact upon psychological health.

The causality between problem debt and psychological health is difficult to establish. The positive relationship between the two might be explained by perception bias or unobserved factors not captured in cross-section analysis. There is also the possibility of a two-way causality between debt and depression: the anxiety and worry caused by the onset of problem debt might lead to declines in individual psychological health, or alternately an individual's psychological health might lead them to incur problem debts through suboptimal financial choices.

This short article summarises empirical findings from the economics literature on the psychological health effects of personal debt, as well as findings from clinical psychology, psychiatry and epidemiology. The article first summarises findings from the medical literature, then discusses the measurement of psychological health in survey data typically used in economic research on the topic and reviews recent studies which aim to understand the direction of causality between personal debt and psychological health.

Evidence from the Medical Literature

Numerous studies in the medical literature show that high levels of personal debt are associated with poor psychological health and related adverse health behaviours. These studies typically use data from the United States of America (USA) or United Kingdom (UK). For recent published literature reviews in medical journals see Fitch et al. (2011) and Richardson et al. (2013). Medical studies are typically based on samples from health service user populations, i.e. patients (Hatcher 1994; Battersby et al. 2006; Abbo et al. 2008), statistical analysis of cross-section survey data (Clark et al. 2012; Jenkins et al. 2009; Meltzer et al. 2011) or psychological autopsy studies following suicides (Chan et al. 2009; Wong et al. 2008, 2010). In general the literature does not attempt to ascertain the direction of causality between personal debt and psychological health.

Stress is the main mechanism through which problem debt impacts upon psychological health. The experience of stress is related to reduced functioning of the immune system and the direct release of stress-related hormones, which impact upon blood pressure and cardiovascular function. Findings from a large literature are consistent with the idea that life experiences which induce stress are causes of physical and mental illness (Goldberger and Breznitz 1993; McEwen 1998a, b; Cooper 2005; Schneiderman et al. 2005).

Measures of psychological health used in studies on personal debt in the medical literature differ according to the research design. Studies based on data from health service user populations (i.e. patients) typically measure psychological health using practitioner evaluations of patient symptoms, attitudes and behaviours using scaling instruments such as the Clinical Interview Schedule Revised, Beck Depression Inventory or Beck Suicide Intent Scale. Psychological autopsy studies use data from interviews with relatives of suicide completers and information from coroner reports. Studies using cross-sections of survey data make use of self-completion survey modules such as the 12-point General Health Questionnaire (or GHQ-12, see below), also used by economists.

Evidence from a meta-analysis in Richardson et al. (2013) reveals that the majority of studies (78.5%) find that high levels of personal debt are positively associated with poor mental health and also suicide completion, drug and alcohol abuse. Among recent psychological autopsy studies, which typically use only small samples with few controls, Chan et al. (2009) found that 23% of suicides are attributable to personal debt problems. However, Richardson et al. (2013) conclude 'The main problem with the current research is that the vast majority of studies are cross-sectional, meaning that causality cannot be established'.

Measurement of Debt and Psychological Health in the Economics Literature

The economics literature emphasises the causality between personal debt and psychological health, but also focuses on the measurement of debt and health in household survey data, which is the typical form of data used in the literature. Large-scale cross-section or panel surveys used for analysis include the British Household Panel Survey (BHPS) for the UK or Panel Study of Income Dynamics (PSID) for the USA. Panel surveys include self-reported data on individual indebtedness, often at the product level which over formal and informal debts, plus self-reported evaluations of whether the individual or household faces repayment problems or is behind on scheduled payments. Some surveys, such as the UK Wealth and Assets survey, also contain data on non-payment, delinquency, arrears and personal insolvency.

Panel surveys contain self-reported measures of psychological health. These include self-reported medical conditions (e.g. ‘anxiety’, ‘depression’) based on the respondent reporting their own medical history or medical conditions for which they have sought help from a medical professional. Surveys also contain psychological health data constructed from self-completion modules such as the GHQ12 in the BHPS or K-6 screening scale in the PSID. As an example, the GHQ-12 comprises 12 questions which ask respondents about their recent experience of poor psychological health, such as feeling anxious, lack of ability of concentrate on everyday activities, sleep behaviours and evaluations of their own self-worth. Responses to these questions form a 0–12 scale rating an individual’s psychological health state. The advantage of these survey instruments is that they are commonly used by public health professionals to diagnose low-level psychiatric disorders in the general population. There is strong evidence that clinical assessments of the severity of psychiatric illness closely correlate with the number of symptoms reported by the GHQ-12 scale (Goldberg 1985).

To illustrate these data, Table 1 (adapted from Gathergood 2012) provides summary data on individual problem debt status and psychological health from the BHPS. The summary data is based on two samples of BHPS respondents. The first is a 1991–2008 sample spanning the whole survey period which covers the head of each household from which summary statistics on housing payments are calculated. The second is a 1995–2008 sample, also covering the head of each household, from which summary statistics on consumer credit are calculated; this later sample is due to the introduction of questions on consumer credit payments from 1995 onwards only.

Among the whole (Column 1) 8% of respondents reported suffering from ‘anxiety, depression or bad nerves’ the average GHQ-12 score was 2.03, indicating that respondents on average reported suffering two of the 12 indicators of poor psychological health. In contrast, 16% of individuals who reported that meeting payments on their housing rent of mortgage schedule had been a problem over the past 12 months (Column 2) were currently suffering anxiety, and the average GHQ among this group was 3.50.

Among those currently two or more months late on housing payments (Column 3), 21% reported suffering anxiety and the average GHQ-12 score was 4.04. These differences in means for the groups in Columns 2 and 3 compared with the whole sample in Column 1 are in both cases statistically significantly different from zero at the 1% level. For consumer credit (Column 4 and 5), those reporting their payments as a ‘heavy burden’ on their finances also showed, on average, higher GHQ-12 scores and higher rates of reporting suffering from anxiety. The

Personal Debt and Psychological Health, Table 1 Debt problems and psychological health, BHPS sample

	1991–2008			1995–2008	
	1. Whole sample	2. Housing payment problems	3. 2+ months late on housing payments	4. Whole sample	5. Consumer credit payments a heavy burden
N	66,664	6499	1541	54,731	8864
Percentage of sample	100 %	9.7 %	2.3 %	100 %	16.2 %
Psychological health					
GHQ-12 Score (0–12)	2.03	3.50	4.04	2.04	2.87
Suffers anxiety = 1	0.08	0.16	0.21	0.09	0.14

difference in means between the two groups is again statistically significant at the 1% level.

By way of comparison, the size of the differences in psychological health among groups shown in Table 1 (with those late on housing payments 13% points more likely to suffering anxiety and showing GHQ-12 score on average 2.1 points higher) are large compared with comparisons between those in employment and unemployment. In the whole data sample unemployed workers are 4% points more likely to suffer anxiety and have GHQ-12 scores on average 1.6 points higher; hence the relationship between debt problems and psychological health is two to three times stronger than the relationship between unemployment and psychological health in these unconditional comparisons. The relationship between unemployment and health has been researched extensively in the economics literature (on which see Ruhm 2000, 2003).

Data Reliability and Perception Bias

Most studies in the economics literature make use of self-reported survey data on debt and psychological health described in the previous section (Bartel and Taubman 1986; Lea et al. 1995; Hamilton et al. 1997; Drentea 2000; Brown et al. 2005; Lenton and Mosley 2008). However, the reliability of self-reported survey debt data may be compromised if an individual's psychological health state influences their reporting of their debts or answers to debt-related questions.

This may be a particular problem for questions which ask about whether an individual's debts are a 'problem' or 'burden' for them. An individual's perception of the severity of their debt problems may be affected by their psychological health state. An individual with poor psychological health might be more, or less, inclined to subjectively report they are struggling with debts compared to an individual with good psychological health in the same financial situation. This 'perception bias' would lead to biased estimates of the impact of debt upon psychological health.

This potential measurement problem is examined by Bridges and Disney (2010). They use a short-panel of UK household survey data (The Family and Children Survey) to model the relationship between debt and psychological health. They find that objective measures of debt problems (such as self-reported values for arrears or late payment on credit) correlate more weakly with subjective evaluations of poor health than subjective measures of debt problems (such as those from questions which ask individuals about their ability to cope with their financial burdens). Their results indicate that an individual's psychological health state impacts on their subjective evaluation of their debt position.

Furthermore, they model the simultaneous relationship between debt and psychological health in a bivariate probit model and find that the relationship between both objective and subjective measures of problem debt and psychological health weakens further. They conclude that poor psychological health affects an individual's perception of their financial situation and that unobserved heterogeneity in the tendency of individuals to report problems with both their health and their finances explains most of the observed correlation between measures of problem debt and measures of psychological health. They state: 'In conclusion, it appears that much of the observed correlation between self-reported psychological well-being on the one hand and financial circumstances on the other, is a person-specific effect, associated with the individual' psychological make-up about which, as economists, we can offer little expertise'.

An alternative empirical approach which avoids potential perception bias is to use health and debt data that is not self-reported but derived from external, verifiable sources. Verifiable sources of individual-level health data include patient medical records and information collected form nurse visits conducted as part of the survey process (nurse visits are used in, for example, the English Longitudinal Survey of Ageing). Verifiable sources of credit and debt records include data from credit reference agencies or lenders. However, it is unlikely that a researcher would be able to combine individual health and debt data from different sources due to data availability and restrictions. An alternative empirical approach using administrative data is to combine local-level credit and debt data with local-level health information, an approach used by Currie and Tekin (2011) discussed below.

Evidence from Panel Data and Instrumental Variable Approaches to Measurement

Descriptive studies based on cross-sections of data provide evidence for an association between problem debt and poor psychological health. The evidence from descriptive studies is strongly suggestive that there is a causal link between problem debt and poor psychological health, but in itself it is not conclusive. Ascertaining the direction of causality between the two involves identifying the direction of causality in a way that rules out the possibility of reverse causality or the existence of omitted factors which co-determine both problem debt and psychological health.

A recent study by Gathergood (2012) exploits the panel dimension of BHPS survey data and uses exogenous instruments for problem debt. Exploiting the panel dimension, Gathergood (2012) shows that a simple analysis of the dynamics of problem debt, GHQ-12 scores and reported rates of anxiety shows that the onset of problem debt is associated with worse existing psychological health. This analysis suggests that poor psychological health is a precursor to problem debt.

Evidence for this is shown in Table 2, reproduced from Gathergood (2012). Table 2 provides summary data on GHQ-12 scores and reported rates of anxiety for different groups in the BHPS sample defined by their problem debt status over two waves of the survey for individuals present in at least two consecutive waves of the BHPS panel. The first two rows of data provide average values for two groups of individuals: those who do not have housing payment problems in the first wave (T) but report housing payment problems in the second wave (T + 1); and those who do not have housing payment problems in either wave.

As can be seen from the table, at the time of the second wave (T + 1) the difference in GHQ-12 scores and rates of anxiety among these groups is large at 1.60 points on the GHQ-12 score and 9% points difference in rates of anxiety. However,

Personal Debt and Psychological Health, Table 2 Transition matrix: entry into debt problems by psychological health measures

	GHQ-12 score		Anxiety-related illness	
	Mean (S.D.) at	Mean (S.D.) at *t* + 1	% (S.D.) at *t*	% (S.D) at *t* + 1
Housing payments				
No payment problems at T, payment problems at T + 1	2.97	3.4	0.14	0.16
(*N* = 2413)	(3.59)	(3.87)	(0.35)	(0.37)
No payment problems at T, no payment problems at T + 1	1.78	1.80	0.07	0.07
(*N* = 42,134)	(2.90)	(2.95)	(0.26)	(0.26)
Not 2+ months late at T, 2 + months late at T + 1	3.48	4.06	0.19	0.24
(*N* = 648)	(3.88)	(4.16)	(0.39)	(0.43)
Not 2+ months late at T, Not 2 + months late at T + 1	1.93	1.94	0.08	0.08
(*N* = 43899)	(3.03)	(3.07)	(0.27)	(0.27)
Consumer credit repayments				
Not a heavy burden at T, Heavy burden at T + 1	2.42	2.64	0.12	0.12
(*N* = 3561)	(3.37)	(3.53)	(0.32)	(0.33)
Not a heavy burden at T, Not a heavy burden at T + 1	1.78	1.78	0.08	0.08
(*N* = 31,949)	(2.94)	(2.96)	(0.27)	(0.27)

those who reported housing payment problems at the time of the second wave had, on average, much higher GHQ-12 scores and rates of anxiety at the time of the first wave (T), at which they did not report housing payment problems. The increase in GHQ-12 score between waves for this group is only 0.43 points and the increase in the rate of suffering of anxiety among this group is only 2% points.

This evidence suggests that most of the difference in psychological health between those with and without housing payment problems is due to the pre-existing level of psychological health. This pattern is repeated for the other debt problem categories shown in Table 2, with in each case the existing levels of psychological health being much worse among those experiencing the onset of being at least 2 months behind on housing payments or facing a ‘heavy burden’ of their consumer credit debts, compared with existing levels of psychological health among those not experiencing the onset of payment problems. These results show that selection into debt problems on the basis of poor psychological health explains most of the observed association between problem debt and psychological health in the cross-section comparison.

To test whether perception bias compromises the reliability of the self-reported data, Gathergood (2012) uses two instruments for the self-reported problem debt responses in the BHPS questionnaire. Firstly, self-reported debt problems are instrumented using lender-provided measures of local-level delinquency rates on consumer credit and mortgage debt. These local-level measures correlate with geographic variation in reported rates of problem debt, but would not correlate with the variation in purely perceived debt problems induced by psychological health status.

Secondly, Gathergood (2012) examines the relationship between the self-reported data provided by the respondent in the survey (which in all cases is the head of household) and the psychological health of his or her partner or spouse. If the head of household’s perception of a payment difficulty arises due to his or her mental health state and not due to an actual difficulty, we would not expect to find a positive relationship between the head of household’s answers to the payment difficulty questions and the psychological health of the household head’s spouse or partner.

Results from both approaches yield estimates of the impact on problem debt on psychological health which are very similar in magnitude and statistical significance to those returned from estimates not using these instrumental variables strategies. As a result, Gathergood (2012) concludes that perception bias is not a driver of the observed variation in psychological health in the BHPS survey data.

Identification of Causal Effects

This final section reviews existing evidence on the causal relationship between problem debt and mental health. Studies based on household panel data have shown that the relationship is not explained by unobserved individual specific drivers of both debt and mental health or short-term changes in economic circumstance such as income reductions, job loss or other adverse events. Brown et al. (2005) conducted a panel data analysis using the BHPS and an individual fixed effects modelling approach, estimating the impact of changes in personal debt on mental health using within-individual variation over time. They controlled for a broad range of covariates, including income, employment, respondent physical health status, and physical and mental health status of other family members, plus additional socio-economic controls. Controlling for these they found a statistically significant and positive association between problem consumer credit debt and poor mental health. However, their results do not identify the direction of causality.

Two recent approaches to estimating the causal relationship between personal debt and mental health are those adopted by Gathergood (2012) and Currie and Tekin (2011). Both are based on sources of exogenous variation in financial circumstances derived from the housing market and have arisen in light of house price booms and busts in the USA and the UK. The key requirement of an instrument for problem debt is that an instrument is correlated with individual-level experience of problem debt, but is exogenous to individual changes in psychological health.

The first approach, used in Gathergood (2012), exploits movement in house prices as exogenous changes in the financial situation of individual households. Gathergood (2012) argues that local-level shocks to house prices are a source of exogenous variation in home equity and that these shocks impact upon the severity of payment problems. The reasoning behind this approach is that if a household defaults on its housing debts, it is unambiguously better for a household to face default with more rather than less housing equity. Defaulting with negative housing equity implies foreclosure, sale of the property and a resulting unsecured consumer debt which may be pursued by the credit through bankruptcy. Defaulting with positive housing equity at least allows the household to sell the home and pay down outstanding debts, or potentially renegotiate the mortgage contract and extract equity if the reason for default is temporary, such as a short-term income shock.

Using local-level house price movements as an exogenous source of home equity shock, Gathergood (2012) showed that a negative home equity shock for an individual with mortgage payment problems leads to a worsening in psychological health, whereas a positive home equity shock leads to an improvement in psychological health. Estimates also show that home equity changes in and of themselves do not impact upon psychological health directly. These results are shown for both the GHQ-12 measure of psychological health and the self-reported anxiety measure of psychological health.

Finally, Gathergood (2012) also showed that there is a social norm dimension to the relationship between problem debt and psychological health. Results show that individuals experiencing the onset of problem debt (either mortgage debt or consumer credit debt) in localities in which the bankruptcy of repossession (foreclosure) rate is higher suffer less deterioration in psychological health compared with individuals who experience the onset of problem debt in localities in which the bankruptcy or repossession rate is lower. This result also holds for both measures of psychological health.

An alternative approach to identifying the causal effect of problem debt is that adopted by Currie and Tekin (2011). They exploited the national wide-scale nature of the housing foreclosure crisis in the USA from 2005 as an unanticipated shock to household debt problems not caused by individual health. While some foreclosures are due to individual health deterioration and related job loss or medical expenses, the nature of the US housing crises was of such a scale that it was not conceivably caused by a widespread outbreak of ill health.

They used data on foreclosures at the zip code level in four US states (Arizona, California, Florida and New Jersey), combining this administrative data with data on Emergency Room (ER) visits and hospitalisations. The four states used in the analysis saw 50% of foreclosures in 2008. They investigate whether increases in foreclosures at the zip code level are linked to higher rates of ER admissions and hospitalisations.

As there is much geographic variation between neighbourhoods which might be linked both with foreclosure and psychological health (such as rates of poverty, unemployment and other socio-economic characteristics), the authors identify causal effects using within-zip code variation over time. The econometric model estimated to determine causal effects includes controls for variation in healthcare quality across localities and over time which might impact upon the decision to seek medical treatment when suffering adverse medical effects related to foreclosure.

Results show that an additional 20 foreclosures in a locality lead to a 2.8% increase in non-elective hospital visits to local hospital facilities, either as ER visits or as hospital admissions. Results also show a lag from foreclosure events to hospital visits, with estimates implying that an additional 80 foreclosures over the course of the prior four quarters lead to a 0.94% increase in the number of hospital visits. Currie and Tekin (2011) show that the adverse health effects of foreclosure are not evenly distributed across household types. Foreclosures are particularly associated with increases in non-elective visits involving young children, potentially reflecting young working households with children as a group particularly susceptible to foreclosure. They also show that house price movements not associated with foreclosure have no statistically significant impact upon hospital admissions, suggesting no severe health effects arise from movements in house prices alone (a result which mirrors the finding in Gathergood (2012) that house prices movements only impact upon psychological health for mortgage holders with existing mortgage payment problems).

Directions for Further Research

Directions of for further research include using alternative instruments for problem debt to identify the impact of exogenous problem debt events on individual health. It is possible that the nature and extent of the health impact is dependent upon the type of problem debt, in particular whether the problem debt relates to short-term debts without lasting impacts or long-term debt problems (for example those associated with poverty). It may be the case that individuals with perpetual debt problems due to poverty and low income adapt to the psychological experience of problem debt such that the impacts are less severe.

An alternative direction for future research is also to investigate the impact of poor psychological health on debt decisions. The approaches to causality used in studies in economics to date have focused on estimating the causal impact of problem debt on health while ruling out reverse causality. It may be the case that there are causal links from poor psychological health to problem debt and studies of decision making among those with poor psychological health might also consider other aspects of financial management such as budgeting and retirement saving.

See Also

- ▶ Health Econometrics
- ▶ Health Economics
- ▶ Identification
- ▶ Partial Identification in Econometrics
- ▶ Subprime Mortgage Crisis

Selected Works

Bridges, S., and R. Disney. 2010. Debt and depression. *Journal of Health Economics* 29: 388–403.

Gathergood, J. 2012. Debt and depression: Causal links and social norm effects. *Economic Journal* 122:1094–14.

Richardson, T., P. Elliott, and R. Roberts. 2013. The relationship between personal unsecured debt and mental physical health: A systematic review and meta-analysis. *Clinical Psychology Review*, forthcoming.

Bibliography

Abbo, C., S. Ekblad, P. Waako, E. Okello, W. Muhwezi, and S. Musisi. 2008. Psychological distress and associated factors among the attendees of traditional healing practices in Jinja and Iganga districts, Eastern Uganda: A cross-sectional study. *International Journal of Mental Health Systems* 2: 16.

Bartel, A., and P. Taubman. 1986. Some economic and demographic consequences of mental illness. *Journal of Labor Economics* 4: 243–256.

Battersby, M., B. Tolchard, M. Scurrah, and L. Thomas. 2006. Suicide ideation and behaviour in people with pathological gambling attending a treatment service. *International Journal of Mental Health and Addiction* 4: 233–246.

Bridges, S., and R. Disney. 2010. Debt and depression. *Journal of Health Economics* 29: 388–403.

Brown, S., K. Taylor, and S. Wheatley-Price. 2005. Debt and distress: Evaluating the psychological cost of credit. *Journal of Economic Psychology* 26: 642–663.

Chan, S.S., H.F. Chiu, E.Y. Chen, W.S. Chan, P.W. Wong, C.L. Chan, Y.W. Law, and P.S. Yip. 2009. Population-attributable risk of suicide conferred by axis I psychiatric diagnoses in a Hong Kong Chinese population. *Psychiatric Services* 60: 1135–1138.

Clark, C., C. Pike, S. McManus, J. Harris, P. Bebbington, T. Brugha, R. Jenkins, H. Meltzer, S. Weich, and S. Stansfeld. 2012. The contribution of work and non-work stressors to common mental disorders in the 2007 Adult Psychiatric Morbidity Survey. *Psychological Medicine* 42: 829–842.

Cooper, C.L. 2005. *Handbook of stress medicine and health*. Boca Raton: CRC Press.

Currie, J., and E. Tekin. 2011. *Is there a link between foreclosure and health?* NBER working paper 17310. Cambridge, MA: National Bureau of Economic Research.

Drentea, P. 2000. Age, debt and anxiety. *Journal of Health and Social Behaviour* 41: 437–450.

Fitch, C., S. Hamilton, P. Bassett, and R. Davey. 2011. The relationship between personal debt and mental health: A systematic review. *Mental Health Review Journal* 16: 153–166.

Gathergood, J. 2012. Debt and depression: Causal links and social norm effects. *Economic Journal* 122: 1094–1114.

Goldberg, D. 1985. Identifying psychiatric illnesses among general medical patients. *British Medical Journal* 291: 161–162.

Goldberger, L., and S. Breznitz (eds.). 1993. *Handbook of stress: Theoretical and clinical aspects*. New York: Free Press.

Hamilton, V., P. Merrigan, and E. Dufresne. 1997. Down and out: Estimating the relationship between mental health and unemployment. *Health Economics* 6: 397–406.

Hatcher, S. 1994. Debt and deliberate self-poisoning. *British Journal of Psychiatry* 164: 111–114.

Jenkins, R., P. Bebbington, T. Brugha, D. Bhugra, M. Farrell, J. Coid, N. Singleton, and H. Meltzer. 2009. Mental disorder in people with debt in the general population. *Public Health Medicine* 6: 88–92.

Lea, S., P. Webley, and C. Walker. 1995. Psychological factors in consumer debt: Money management, economic socialisation and credit use. *Journal of Economic Psychology* 16: 681–701.

Lenton, P., and P. Mosley. 2008. Debt and health. In: *Sheffield economic research paper series* Number 2008004, University of Sheffield.

McEwen, B.S. 1998a. Stress, adaptation, and disease: Allostasis and allostatic load. *Annals of the New York Academy of Sciences* 840: 33–44.

McEwen, B.S. 1998b. Protective and damaging effects of stress mediators. *New England Journal of Medicine* 338(3): 171–179.

Meltzer, H., P. Bebbington, T. Brugha, R. Jenkins, S. McManus, and M.S. Dennis. 2011. Personal debt and suicidal ideation. *Psychological Medicine* 41: 771–778.

Richardson, T., P. Elliott, and R. Roberts. 2013. The relationship between personal unsecured debt and mental physical health: A systematic review and meta-analysis. *Clinical Psychology Review* 33: 1148–1162.

Ruhm, C.J. 2000. Are recessions good for your health? *Quarterly Journal of Economics* 115: 617–650.

Ruhm, C.J. 2003. Good times make you sick. *Journal of Health Economics* 22: 637–658.

Schneiderman, N., G. Ironson, and S.D. Siegel. 2005. Stress and health: Psychological, behavioral, and biological determinants. *Annual Review of Clinical Psychology* 1: 607.

Wong, P.W.C., W.S.C. Chan, E.Y.H. Chen, S.S.M. Chan, Y.W. Law, and P.S.F. Yip. 2008. Suicide among adults aged 30–49: A psychological autopsy study in Hong Kong. *BMC Public Health* 8: 147.

Wong, P.W., W.S. Chan, Y. Conwell, K.R. Conner, and P.S. Yip. 2010. A psychological autopsy study of pathological gamblers who died by suicide. *Journal of Affective Disorders* 120: 213–216.

Personnel Economics

Edward Lazear

Keywords

Agency problems; Human resources management; Incentive theory; Labour economics; Personnel economics; Prospect theory; Risk aversion; Stock options; Tournament theory

JEL Classifications

M50

Personnel economics is the application of economic and mathematical approaches and econometric and statistical methods to traditional questions in human resources management. Many of the issues studied by personnel economists can be found in traditional textbooks written by organizational behaviour scholars and other human resources specialists. Economists have something new to say about these issues, however, primarily because economics provides a rigorous, and in many cases more straightforward, way to think about these human resources questions than do the more sociological and psychological approaches. Certain questions, especially those dealing with compensation, turnover and incentives, are inherently economic. Others, like those associated with non-monetary aspects of the job, norms, teamwork, worker empowerment and peer relationships, while seemingly non-economic, are capable of being informed by economic reasoning. Economists have the advantage of knowing how to strip away extraneous detail and focus on the essentials. This allows them to provide precise and reasoned answers that are testable and refutable and thereby follow the scientific method used by the physical sciences. One drawback of the economic approach, when applied to human resources (and other) issues, is that sometimes its simplifications miss some of the descriptive detail that gives depth and understanding to a situation.

What are the main goals of personnel economics? The primary goal is to provide positive analysis of human resources practices and methods. When do firms choose to use one form of compensation over another? When are teams important? When is job rotation effective? When are certain benefits or stock grants given to workers? The list extends. But in addition to being able to describe what is, personnel economics is more normative than most fields of economics. Perhaps because the subject was taken up by business school economists whose job is to teach managers what to do, personnel economics has not shied away from being somewhat prescriptive. In part, personnel economics is an attempt to look inside the black box. It is an imperialistic attempt by economists to do what Alfred Marshall (1890) said that 'economists do not do': Marshall's famous statement that it is not the economist's business to tell the brewer how to brew beer has not been adhered to when it comes to personnel economics. Personnel economists often attempt to do precisely that; namely, to use the tools of economics to understand and sometimes even to guide practitioners and consultants in their trade.

From a practical point of view, personnel economics is important. Labour accounts for approximately 70 per cent of costs and this number has been reasonably stable over time. Changes that affect labour productivity, turnover, or aspects of compensation can have quite dramatic effects on company profits. In one recent example (see Lazear 2000b), a company altered its method of pay and consequently experienced a 44 per cent increase in productivity in a period of about six months. Such large shifts in productivity are extremely rare and come about mostly with major innovations in technology. Although changes of this magnitude are likely to be unusual even in the realm of personnel economics, the point remains that action on the cost front is likely to involve labour issues because labour is the primary component of cost for most firms.

Personnel, which has become more fashionably known as human resources management, has been around as an academic and practical subject for at least the last 50 years. But personnel economics takes a different view of many of the same questions and issues that are part of standard human resources management. How does personnel economics differ from 'old-style' personnel analyses? Primarily, the difference lies in the rigour associated with the economic approach, which is absent from traditional analyses. Personnel economics is, above all, economics. As such, it follows the approach used by economists. This approach is described in Lazear (2000a) and again in Lazear (2000c). Much of the material in the next few paragraphs is taken directly from Lazear (2000c).

First, personnel economics assumes that the worker and firms are rational maximizing agents. Constrained maximization is the basic building block of all theories in personnel economics. Empirical analyses focus on tests of rational, maximizing models. When evidence contradicts a model, the approach of personnel economists is to think more carefully about the nature of the model set-up, rather than to drop the assumption of rationality. The assumption of maximizing rational behaviour in personnel economics is in large part done in order to allow the analyst to express complicated concepts in relatively simple, albeit abstract, terms.

In many respects, this is the main virtue of personnel economics. The typical human resources text eschews generalization, arguing that each situation is different. The economist's approach is the opposite, following the scientific method that places a premium on discovering the underlying general principle.

A second distinguishing feature is that personnel economists focus on equilibrium. Like the physical sciences, almost all theories in personnel economics are consistent with some notion of equilibrium. This differs dramatically from the approaches used in other social sciences, primarily psychology and sociology. Psychologists are interested in individual behaviour and so equilibrium at the market level is not central. But when discussing issues at the level of the firm, especially those that are imbedded in a market context, equilibrium is essential. Personnel economics differs from other approaches to studying personnel in that, as in all branches of economics, there is no free lunch. Firms hire workers in a competitive labour market and cannot simply take advantage of them. Workers cannot be induced to do things that they do not want to do without appropriate compensation, either in the form of money or some other non-monetary reward.

Consider, for example, the provision of incentives. A psychologist might argue that a particular compensation structure offers stronger incentives than another – the best known is Kahneman and Tversky's (1979) prospect theory, which argues that losses impose more disutility than an equivalent gain produces utility. This implies that penalties are more powerful incentive providers than are bonuses – and might suggest that, as a result, firms should adopt the more powerful form of compensation. This ignores the fact that effort is costly and in equilibrium firms that induce more effort must pay higher wages. It is possible that too much effort results because the additional output from the effort may be smaller than the additional amount necessary to compensate the worker for the increased effort.

Third, efficiency is a central concept of personnel economics. Adam Smith's early notion of the invisible hand makes its way into personnel economics. Individuals who maximize their own utility and interact with firms that maximize profits generate behaviour that usually makes both parties better off. When efficiency suffers, say as a result of moral hazard problems that arise in the agency literature, the economist pushes the analysis to another level, asking what actions might firms and/or workers take to alleviate such inefficiency. Taking this further step assists in making better positive predictions and also normative prescriptions for the business student.

In an analogous vein, personnel economists think in terms of substitution, where other human resources specialists do not. For example, most firms have a benefits department that is distinct from the compensation department and compensation is defined specifically to include monetary remuneration only. There is no explicit recognition of trade-offs, and non-economists frequently think in terms of providing some market level of each job attribute rather than thinking in terms of a total package that guarantees some reservation utility.

Some Basic Theory

Much of the early work in personnel economics was on the theory of compensation. This was a natural outgrowth of the agency literature that dates back to 1950 (see Johnson 1950, and later Cheung 1969). The early modern treatments of the agency problem are found in Ross (1973), who lays out the fundamental agency analysis and later Stiglitz (1975) and Bergson (1978). The basic idea in this early work is that the owner and the worker are not the same individual, and so their interests may not be aligned. In particular, the worker wants money, but does not like to put forth effort. The owner wants output, but would rather not pay for it.

The standard agency problem is solved by a piece-rate compensation scheme. In the simple, risk-neutral worker case, the optimal scheme pays the worker the full value of his output on the margin, setting the piece rate equal to the (net) value of output. Generally coupled with this is a rent-sharing parameter so that

$$\text{Wage} = a + b\,q, \tag{1}$$

where q is output, $b = 1$ and a is set so that the worker is just indifferent between taking the job and taking his next best alternative.

The analysis becomes more complicated, but not fundamentally different, when noise in production and risk aversion are introduced. The most complete early analysis of this is contained in Hölmstrom (1979). The primary result is that there is now a trade-off. Because workers do not like risk, the firm must dampen the relation of wages to q. There is a trade-off between insurance and incentives. In the context of (1), full insurance can be provided by setting $b = 0$, but as a result, the worker has no incentive to put forth. Were $b = 1$, incentives are provided but the worker bears the full risk. The solution, which generally uses a nonlinear compensation scheme, forces the worker to bear some risk and sacrifices effort relative to the risk-neutral case. Another variant on this scheme is presented by Gibbons (1987). Gibbons considers the case where only the worker knows the difficulty of the job and only the worker knows his true action. Under these circumstances, Gibbons shows that workers will restrict output.

Although piece-rate incentive pay characterizes part of the labour market, especially those jobs where output is easily measured, other jobs, perhaps most, do not lend themselves to piece-rate pay. In such cases firms pay salaries, defined as pay based on an input measure, like hours of work, rather than an output measure, like sales. Lazear (1986) describes the factors that lead firms to choose between paying on output or paying on input. But salaries are not fixed and motivation is provided to workers by altering salaries over time based on performance. When absolute output is difficult to observe, workers are ranked, one relative to another, and promotions that are awarded to the better workers serve as motivation. This logic forms the basis of tournament theory, which shows that a well-designed promotion scheme based on rank alone is a perfect substitute for a piece-rate scheme. See Lazear and Rosen (1981).

There are three basic principles of tournament theory. First, prizes are fixed in advance and depend on relative rather than absolute performance. Second, larger spreads in wages at different levels of the hierarchy motivate those at lower levels to put forth more effort. Third, there is an optimal spread. Although a greater spread increases effort, at some point the additional wages necessary to compensate workers for the increased effort is larger than the additional output generated. The formal analysis sets up a problem in which workers maximize

$$\underset{\mu_j}{\text{Max}} \quad W_1 P + W_2(1-P) - C\left(\mu_j\right) \tag{2}$$

where W_1 is the wage to the winner who gets the promotion, W_2 is the wage to the loser who is not promoted, P is the probability that a worker gets promoted to the high paying job by out-performing his rival, and μ_j is effort, having cost $C(\mu_j)$. The first-order condition to the worker's problem is

$$(W_1 - W_2)\frac{\partial P}{\partial \mu_j} - C'\left(\mu_j\right) = 0 \tag{3}$$

A firm takes (3) into account and sets wages, W_1 and W_2, so as to maximize profits subject to paying enough on average to attract workers to the job.

Some implications follow from (3). First, an increase in $W_1 - W_2$ implies a higher equilibrium level of effort, since $C'(\mu_\mathrm{j})$ is increasing in μ. Larger rises associated with promotion increase the equilibrium level of effort. If promotion is valuable, workers work hard to obtain a promotion.

Second, a decrease in $\partial P/\partial \mu_j$ lowers effort. It is straightforward to show that an increase in noise or luck lowers $\partial P/\partial \mu_j$. Volatile industrial environments generally have highly skewed earnings, which serve to offset the tendency for workers to give up when there is too much randomness associated with the promotion decision.

Additional implications follow. Because nepotism reduces the effect of effort on changing the probability of winning, nepotism kills off effort in an organization. Additionally, if too many workers are competing for a given promotion, incentives are weak because effort does not alter the probability of winning very much. This provides a rationale for limiting the competition in a promotion race.

Some workers will never be promoted again and know it, but it may nevertheless be important to keep them motivated. Upward-sloping experience-earnings profiles that result in back-loaded compensation provide incentives. Workers are paid less than they are worth in the early years of their job, but more than they are worth when senior. The higher-than-alternative wage that they receive in the latter years keeps them performing on the job because they do not want to lose the (*ex post*) rents associated with satisfactory performance on the current job. To clear the market, they accept lower wages when young so that over their working life, wages add up to their productivity. Unlike the efficiency wage literature (see Shapiro and Stiglitz 1984; and Akerlof 1984 for the classic reference) that focuses on how unemployment can emerge, the thrust in personnel economics has been to ask whether less constrained compensation schemes can remove the excess supply of labour.

The theoretical literature in personnel economics is now quite rich. Topics of hiring and firing, the trade-off between money and benefits, evaluation and worker empowerment and delegation of authority are only a few of the topics analysed.

Empirical Literature

Theory is most valuable when it provides predictions that can be verified or refuted by real-world experiences. Personnel economics has an array of implications and some have been analysed in the context of data from businesses that try different aspects of personnel and compensation policy.

There are many examples, but only a few are listed here. Most obvious are tests of incentive theory. Compensation variations provide fertile ground on which to examine worker responses to incentives. A number of recent papers have examined piece-rate pay and its implications for worker behaviour, both in terms of incentives and sorting (see Lazear 2000; Paarsch and Shearer 1999; Fernie and Metcalf 1999; Eriksson and Villeval 2004). The finding is that a move from hourly pay to piece-rate pay generally increases output and attracts a more productive workforce. The incentive models of the theoretical personnel economics literature are excellent predictors of real behaviour. They do not imply that piece-rate pay is superior to hourly wages. Higher output comes at a cost (higher wages, sometimes lower quality) and the choice of compensation scheme depends on the factors described in the theory, such as measurement costs and quality–quantity trade-offs. Other examples that tie pay to output involve evidence on the nature of executive compensation and formulae that link pay to measures of output. In most cases, earnings of top executives are tied to a measure of team performance instead of, or in addition to, individual performance. The metric is stock or bonuses that are based on earnings (the best known is Jensen and Murphy 1990).

Stock and stock options have become an important part of compensation for high-level managers and for knowledge workers in general. Stock may provide incentives, but for most workers the incentive effects of stock ownership must be quite small because they own only a small part of the firm and capture a trivial part of the returns to their effort. Some argued that stock ownership, because of its gradual vesting structure, provides incentives to stay on the job (Oyer 2004; Oyer and Shaefer 2005). Recently, evidence has become available that demonstrates the significant effect of non-vested stock options and certain types of bonus payments in employee retention (Russell 2005). It is important to point out that the fact that non-vested compensation provides incentive effects does not imply that they should be used. Again, this is part of thinking about equilibrium. Inefficient retention provides benefits to the firm that fall short of worker costs and in equilibrium vanish as firms find that they must pay workers too much when they create excessive retention incentives.

There is also empirical support for the tournament view of labour markets. Larger prize spreads induce more effort; wage structures seem consistent with tournament structures, and workers behave selfishly and fail to cooperate when relative performance pay is too strong (see for example Ehrenberg and Bognanno 1990; Drago and Garvey 1998; Eriksson 1999; Falk and Fehr 2006; Knoeber 1989).

Related to tournaments, other empirical evidence provides support that upward-sloping experience-earnings profiles are used to motivate workers. These studies use the implications of the theory with respect to variations in use of the method across demographic groups and job complexity. More complex jobs with harder-to-measure output must seek forms of incentive pay other than pure piece rates. The evidence suggests that steeper profiles are used in jobs where measurement is less straightforward (Hutchens 1986, 1987, 1989). Others have pointed out that long-term employment incentives can only be used for workers who are permanently attached to the labour market. Those who have shorter expected employment duration should be more likely to be paid piece rates; those with permanent attachments should be relatively more likely to see upward-sloping experience-earnings profiles. The evidence supports this claim (see Goldin 1986).

Another kind of evidence relates to how human resource practices affect productivity. (The best known are a series of papers by Ichniowski and Shaw, for example, Ichniowski et al. 1997, 2001, 2007). Compensation is only one way that worker productivity can be altered. The actual organization of work can matter. Working in teams, using job rotation, providing training, sharing information and a number of other practices have been shown to have significant effects on worker productivity. Interestingly, the more modern human resource practices are always coupled with some kind of (team) incentive pay (Jensen and Kevin 1990). Apparently, the practices themselves, without the incentives to use and implement them, do not produce the desired effects on output.

Conclusion

Personnel economics has been among the most active fields in labour economics since the 1980s. There are three reasons. First, the questions it raises are fundamentally important. Labour is the key factor of production and understanding the ways by which labour productivity can be altered is central to the economics of business. Second, there has been an abundance of theoretical insights that are satisfying not only at the intellectual level, but that seem inherently sensible and able to explain the real world. Third, the theories provide specific implications that can be tested: when the analyses are brought into contact with real data, the theories are confirmed.

See Also

- ▶ Efficiency Wages
- ▶ Experimental Labour Economics
- ▶ Labour Economics
- ▶ Labour Economics (New Perspectives)
- ▶ Labour Market Institutions.

Bibliography

Akerlof, G. 1984. Gift exchange and efficiency wage theory: Four views. *American Economic Review* 74: 79–83.

Bergson, A. 1978. Managerial risks and rewards in public enterprises. *Journal of Comparative Economics* 2: 211–225.

Cheung, S.N.S. 1969. *The theory of share tenancy: With special application to Asian agriculture and the first phase of Taiwan land reform*. Chicago: University of Chicago Press.

Drago, R., and G.T. Garvey. 1998. Incentives for helping on the job: Theory and evidence. *Journal of Labor Economics* 16: 1–25.

Ehrenberg, R.G., and M.L. Bognanno. 1990. Do tournaments have incentive effects? *Journal of Political Economy* 98: 1307–1324.

Eriksson, T. 1999. Executive compensation and tournament theory: Empirical tests on Danish data. *Journal of Labor Economics* 17: 262–280.

Eriksson, T., and M.-C. Villeval. 2004. *Other-regarding preferences and performance pay. An experiment on incentives and sorting*. Discussion Paper No. 1191, Institute for the Study of Labor (IZA), Bonn.

Falk, A., and E. Fehr. 2006. *The power and limits of tournament incentives*. Mimeo: University of Bonn.

Fernie, S., and D. Metcalf. 1999. It's not what you pay it's the way that you pay it and that's what gets results: Jockeys' pay and performance. *Labour* 13: 385–411.

Gibbons, R. 1987. Piece-rate incentive schemes. *Journal of Labor Economics* 5: 413–429.

Goldin, C. 1986. Monitoring costs and occupational segregation by sex: A historical analysis. *Journal of Labor Economics* 4: 1–27.

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

Hutchens, R.M. 1986. Delayed payment contracts and a firm's propensity to hire older workers. *Journal of Labor Economics* 4: 439–457.

Hutchens, R.M. 1987. A test of Lazear's theory of delayed payment contracts. *Journal of Labor Economics* 5: S153–S170.

Hutchens, R.M. 1989. Seniority, wages and productivity: A turbulent decade. *Journal of Economic Perspectives* 3(4): 49–64.

Ichniowski, C., K. Shaw, and G. Prennushi. 1997. The effects of human resource management practices on productivity. *American Economic Review* 86: 291–313.

Ichniowski, C., K. Shaw, and W. Boning. 2001. *Opportunity counts: Teams and the effectiveness of production incentives*. Working Paper No. 8306. Cambridge, MA: NBER.

Ichniowski, C., K. Shaw, and A. Bartel. 2007. How does information technology really affect productivity? Plant-level comparisons of product innovation, process improvement and worker skills. *Quarterly Journal of Economics* (forthcoming).

Jensen, M.C., and J.M. Kevin. 1990. Performance pay and top management incentives. *Journal of Political Economy* 98: 225–264.

Johnson, D.G. 1950. Resource allocation under share contracts. *Journal of Political Economy* 58: 111–123.

Kahneman, D., and A. Tversky. 1979. Prospect theory: An analysis of decision under risk. *Econometrica* 47: 263–291.
Knoeber, C.R. 1989. A real game of chicken: Contracts, tournaments, and the production of broilers. *Journal of Law, Economics & Organization* 5: 271–292.
Lazear, E.P. 1986. Salaries and piece rates. *Journal of Business* 59: 405–431.
Lazear, E.P. 2000a. Performance pay and productivity. *American Economic Review* 90: 1346–1361.
Lazear, E.P. 2000b. Economic imperialism. *Quarterly Journal of Economics* 115: 99–146.
Lazear, E.P. 2000c. The future of personnel economics. *Economic Journal* 110: F611–F639.
Lazear, E.P., and S. Rosen. 1981. Rank-order tournaments as optimum labor contracts. *Journal of Political Economy* 89: 841–864.
Marshall, A.P. 1890. *Principles of economics,* 9th ed, with annotations by C.W. Guillebaud. London/New York: Macmillan, for the Royal Economic Society, 1961.
Oyer, P. 2004. Why do firms use incentives that have no incentive effects? *Journal of Finance* 59: 1619–1649.
Oyer, P., and S. Schaefer. 2005. Why do some firms give stock options to all employees: An empirical examination of alternative theories? *Journal of Financial Economics* 76: 99–133.
Paarsch, H.J., and B. Shearer. 1999. The response of worker effort to piece rates. *Journal of Human Resources* 34: 643–667.
Ross, S.A. 1973. The economic theory of agency: The principal's problem. *American Economic Review* 63: 134–139.
Russell, K. 2005. *Deferred compensation and employee retention*. Ph.D. thesis, Department of Economics, Stanford University.
Shapiro, C., and J.E. Stiglitz. 1984. Equilibrium unemployment as a worker discipline device. *American Economic Review* 74: 433–444.
Stiglitz, J.E. 1975. Incentives, risk, and information: Notes toward a theory of hierarchy. *Bell Journal of Economics and Management Science* 6: 552–579.

Persons, Warren Milton (1878–1937)

George S. Tavlas

Keywords

Business-cycle measurement; Correlation coefficient; Equation of exchange; Persons, W. M.; Quantity theory of money; Time series analysis

JEL Classifications

B31

Persons was born on 12 March 1878 in West De Pere, Wisconsin; he died on 11 October 1937 in Cambridge, Massachusetts. After studying mathematics and economics at the University of Wisconsin, he taught at several universities, including Harvard where he became, in 1919, the first editor of the *Review of Economics and Statistics.*

Persons' primary contribution was in the application of statistical methods to the analysis and measurement of economic fluctuations. Early on in his career, he was involved in the debate regarding the empirical validity of the quantity theory of money. He introduced the use of the correlation coefficient into the quantity theory literature as a means of testing the relationships among the variables in the equation of exchange (Persons 1908) and was the first to employ first-differencing in the quantity-theory debate to remove trend from his data (Persons 1910).

At Harvard, Persons set out to put differing numerical series into a form which comparisons could be made both among the various series and between different points of time in a given series. In this regard, he devised the 'Harvard Barometer' technique of eliminating seasonal and trend influences from time series. Comparisons of the timing of the adjusted series showed systematic differences among them, and led Persons to emphasize the short-run, periodic, nature of business fluctuations. Consequently, in *Forecasting Business Cycles* (1931), he predicted an end to the business downturn then under way by March 1931. His prediction of an early end to the Great Depression, combined with his advocacy of fiscal retrenchment to combat the depression, may have served to deflect the profession from his substantial contribution to the literature on business-cycle measurement.

Selected Works

1908. The quantity theory as tested by Kemmerer. *Quarterly Journal of Economics* 22: 274–289.
1910. The correlation of economic statistics. *Quarterly Publications of the American Statistical Association* 12(92): 287–322.

1919. *Indices of general business conditions.* Boston: Harvard University Committee on Economics.
1928. *The construction of index numbers.* Boston: Houghton Mifflin.
1931. *Forecasting business cycles.* New York: Wiley.

Pesch, Heinrich (1854–1926)

H. C. Recktenwald

Abstract
After studies in economics and law at the university of Bonn (1872–5) Pesch became a member of the Society of Jesus in 1876 and took holy orders in 1888. At the University of Berlin (1901–3) he deepened his economic knowledge, which was strongly influenced by the triumvirate Schmoller, Sering and particularly Wagner, who greatly appreciated Pesch's theoretical ability. In the stillness and seclusion of a Berlin cloister his scientific and literary works grew to maturity, among them the five volumes of his *opus magnum Lehrbuch der Nationalökonomie* (1905–23), a treatise of excellent scholarship yet not too proficient in economic analytics. Pesch died in 1926 in Valkenburg, Holland.

After studies in economics and law at the university of Bonn (1872–5) Pesch became a member of the Society of Jesus in 1876 and took holy orders in 1888. At the University of Berlin (1901–3) he deepened his economic knowledge, which was strongly influenced by the triumvirate Schmoller, Sering and particularly Wagner, who greatly appreciated Pesch's theoretical ability. In the stillness and seclusion of a Berlin cloister his scientific and literary works grew to maturity, among them the five volumes of his *opus magnum Lehrbuch der Nationalökonomie* (1905–23), a treatise of excellent scholarship yet not too proficient in economic analytics. Pesch died in 1926 in Valkenburg, Holland.

For his original doctrine of a community, namely the conception of a 'corporative' state which is roughly embodied in the normative programme of the encyclical *Quadragesimo Anno* (*1931*), Pesch is trying to establish norms for a reasonable order and adequate means to realize them. The guiding principle is that man is the origin, bearer and aim of all social life. Like Smith's natural axiom of controlled self-interest (never selfishness), this supreme principle can be rationally determined and empirically tested. Both principles have their roots (and no more) in Aristotle's and Aquinas' ideas. And both are in the final analysis traceable to the Creator or Nature. From this leading idea (*Leitidee*) Pesch derives a social order which stands on three pillars: the subsidiary principle (i.e. the human being and his family rank first, all other groups have to follow and to serve him); the principle of solidarity (the corollary of the *homo socialis*, who is embodied in a community which, however, can never be considered a goal in itself but has to serve man's evolution and unfolding (*Selbstentfaltung*); and the principle of unit (*Ganzheit*) (man is to consider the welfare of his brethren; that is, Smith's fellow-feeling).

On these ethical foundations Pesch explains the economic, social and political order. His eminent ability to arrive at a creative synthesis and his productive intuition in this field are widely underrated in the literature. His concept of order comes close to Smith's 'obvious and simple system of natural liberty'.

Pesch did not postulate absolute free trade and radical social policy (*Sozialpolitik*). He neglected to base his concept on a thorough analysis which would lead him to provable or falsifiable hypotheses. However, he did not argue on the level of most ethical discussions, which up to now have taken place in an illusory world of speculative possibilities and moral rigour rather than in a world constrained by fact and of explanatory hypotheses. Furthermore, Pesch avoided postulating the *absolute* observance of ethical norms by the imperfect man in his natural and imperfect surroundings.

Pesch's theory of order is still more clearly presented in his *Liberalismus, Sozialismus und christliche Gesellschaftsordnung* (1896–9). His intellectual disciple O. von Nell-Breuning, S.J., further developed and refined this theory in his *The Reorganization of the Social Economy* (English trans., 1936).

Indeed, Pesch's concept may substantially contribute (1) to modern discussion on further development of the classical order theory to solve the central problem of how to reconcile both self-interest and group-interest with the *bonum commune* of the society in a world of scarcity (as the theories of public choice or of bureaucracy are trying to do); and (2) to an understanding of the order concept behind the 'theology of liberation'.

Selected Works

1896. *Liberalismus, Sozialismus und christliche Gesellschaftsordnung*. 2nd edn, Freiburg: Herder.

1905–23. *Lehrbuch der Nationalökonomie*. 5 vols. 2nd-4th edns, Freiburg: Herder.

1918. *Ethik und Volkswirtschaft*. Freiburg: Herder.

1924. *Die Volkswirtschaftslehre der Gegenwart in Selbstdarstellungen*. vol. I, ed. F. Meiner. Leipzig: Meiner.

Bibliography

Nell-Breuning, O. von. 1936. *The reorganization of the social economy*. New York: Bruce.

Nell-Breuning, O. von. 1985. *Gerechtigkeit und Freiheit*. Munich: Olzog.

Recktenwald, H.C. 1985. *Ethik, Wirtschaft und Staat: Adam Smiths Politische Ökonomie Heute*. Darmstadt: Wissenschaftliche Buchgesellschaft (See particularly introduction and articles by Friedman, Samuelson, Skinner and Stigler).

Recktenwald, H.C. 1986. *Zur Theorie der ethischen Gefühle*. Darmstadt/Düsseldorf: Wirtschaft und Finanzen.

Samuelson, P.A. 1982. Schumpeter and Marx. In *Schumpeterian economics*, ed. H. Frisch. Eastbourne/New York: Praeger.

Schumpeter, J.A. 1955. *Captialism, socialism and democracy*. 2nd ed. London: Allen & Unwin.

Peso Problem

Karen K. Lewis

Abstract

If market participants expect a future discrete change in asset fundamentals, then rational forecast errors may be correlated with current information and have a mean different from zero in finite samples. This statement may seem inconsistent with the standard assumption that forecast errors are orthogonal to current information and have a mean of zero. By contrast, this article describes how this phenomenon may be rational using the example of the Mexican peso market in which it was first noted. It then illustrates how the peso problem applies more generally to a wide range of asset prices.

Keywords

Efficient markets hypothesis; Foreign exchange risk premium; Friedman, M.; German hyperinflation; Learning; Martingales; Peso problem; Rational expectations; Regimeswitching models; Risk neutrality; Stock price volatility; Term premium; White noise

JEL Classifications

F3

Asset prices are determined by expectations about the paths of future economic variables. Therefore, anticipated discrete changes in the distribution of these variables directly affect asset price behaviour. The 'peso problem' focuses upon how asset prices behave when market traders have expectations about infrequent discrete shifts in economic determinants. With these expectations, the discrete switches can induce behaviour in asset prices that apparently contradicts conventional rational expectations assumptions. The fundamental shifts are rare events and typically occur infrequently, even in relatively large samples. As such, the term ‘peso problem’ is interchangeably with the small-sample inference problems arising from these expected events.

The specific currency reference used in the term ‘peso problem’ may seem at odds with its general potential effects on asset prices. The origins of the term therefore deserve further explanation. The phenomenon is called the ‘peso problem’ because it was first noted in the Mexican peso market. The original source of the term is unknown, though some economists have attributed it to Milton Friedman.

The empirical phenomenon was originally mentioned in writing in the dissertation by Rogoff (1977, 1980) and in publication form by Krasker (1980). Based upon evidence from the Mexican peso futures market from June 1974 to June 1976, Rogoff used the relationship between futures contracts and spot contracts to test market efficiency under rational expectations and risk neutrality. He found that the implications of market efficiency were rejected, but that the behaviour of futures contracts could be explained by the market’s persistent belief that the Mexican peso might be devalued. Consistent with this explanation, the peso was devalued in August 1976.

The Peso Problem in the Mexican Currency Crisis

To illustrate the effects upon asset prices during this period, consider the relationship between the spot and forward rate of a contract for future delivery. If we define S_{t+1} as the logarithm of the future spot rate (dollars per peso) at date $t+1$ and F_t as the logarithm of the forward rate contracted at date t for delivery at date $t+1$, the relationship between the two variables may be written

$$S_{t+1} - F_t = r_t + u_{t+1} \qquad (1)$$

where r_t is the risk premium, the forecast error on the spot rate is $u_{t+1} \equiv S_{t+1} - E_t S_{t+1}$, and E_t is the expectations operator conditional on information available at time t. Through covered interest parity, the difference between the spot and forward rate also equals the return on holding peso deposits over the same period and converting the proceeds back into dollars at date $t+1$. In order to focus on the effect of expectations, the analysis below will ignore the risk premium effect. This assumption is not necessary, however, and much of the literature described below includes models of the risk premium term, r_t.

From April 1954 to August 1976, the spot peso exchange rate was fixed at 0.08 dollars per peso. During this period, which covered over 20 years, the exchange rate was constant. If we use the notation above, therefore, S_{t+1} was equal to a constant, call it S^0. Nevertheless, futures and forward contracts sold at a discount for much of the early 1970s. For example, the year ahead contract on June 1975 and June 1976 futures contracts sold at a discount of 2.6 and 2.7 per cent respectively. Similarly, Mexican peso deposit rates traded higher than dollar deposit rates over this period, implying a forward rate in (1) that was less than the *ex post* spot rate. Therefore, the *ex post* rate of return on holding Mexican peso accounts, $S^0 - F_t$, was systematically positive. Under risk neutrality, this behaviour contradicts the assumption of rational expectations since it implies that the market’s forecast errors, $S_{t+1} - E_t S_{t+1}$, were biased and serially correlated.

At the end of this period, on 31 August 1976, the authorities allowed the Mexican peso to float. Subsequently, the peso fell to 0.05 dollars per peso, implying a devaluation of about 46 per cent. If we define the logarithm of the spot rate associated with this level as S^1, the implied forecast error over this event was $S^1 - F_t \cong -46$ per cent. If one takes account of this large negative observation together with the many small positive observations over the early 1970s the implication is an average forecast error close to zero, which explains the apparent Mexican peso paradox.

Examining how traders with rational expectations would have formed their forecasts helps to define the peso market phenomenon further. Lizondo (1983) postulated that the expected future peso exchange rate could be written as:

$$E_t S_{t+1} = (1 - p_t) S^0 + p_t S^1 \qquad (2)$$

where p_t is the market's assessed probability that the authorities will devalue the peso to S^1 during the next period. Therefore, as long as the peso remains fixed at S^0, the forecast error is

$$u_{t+1} = S^0 - E_t S_{t+1} = p_t\left(S^0 - S^1\right) \qquad (3)$$

Since the Mexican spot rate over the early period was greater than the devalued August 1976 rate, the initial spot rate S^0 was greater than the anticipated rate if devaluation were to occur, S^1. As such, *ex post* forecast errors were systematically positive. The ex *post* bias observed in forecast errors depended upon both the probability of the devaluation, p_t, and the expected size of the fall in the exchange rate, $S^0 - S^1$. On the other hand, for the period when the devaluation occurred, the forecast error was a large negative number, $(1 - p_t)(S^1 - S^0)$.

In a sample with many observations of similar devaluations, forecast errors would be persistently positive with infrequent large negative observations. The frequent small positive forecast errors and the infrequent large negative forecast errors will tend to cancel each other out. Over a sufficiently large sample with enough of the rare events, the forecast errors would roughly sum to zero, as implied by rational expectations. However, the market would appear to make systematic forecast errors between the episodes of discrete changes, even though the forecasts will be unbiased in sufficiently large samples. Even in large samples, therefore, rational forecast errors with a 'peso problem' may be serially correlated.

The Peso Problem in General Asset Prices

Although first noted in the period of the fixed Mexican peso rate, this phenomenon can be found in any forward-looking asset price when market traders anticipate a discrete change in the distribution of its economic determinants. A simple example serves to illustrate the peso problem in general. Suppose that agents rationally anticipate a switch in the process of an economic variable from its current process, R^0, to an alternative, R^1. In this case, rational forecasts of asset prices that depend upon this variable include forecasts of the price conditional upon each regime process. Denote the general asset price as S_t to preserve the same notation as above. Then the expected future value of the asset price is:

$$E_t S_{t+1} = (1 - p_t) E_t\left(S_{t+1} \middle| R^0\right) + p_t E_t\left(S_{t+1} \middle| R^1\right) \qquad (2')$$

where p_t is the market's assessed probability conditional upon time t information that the process will switch to process 1; and where $E_t(S_{t+1}\,|R^i)$ for $i = 0, 1$ is the expected value conditioned upon time t information and upon process i generating the asset's determining variables.

A few examples of peso problem studies serve to illustrate the breadth of its application in diverse settings. Salant and Henderson (1978) considered the effects upon the price of gold from the market's assessed probability that governments might sell their gold holdings in large discrete amounts. In this case, the spot rate S_t represents the price of gold, $E_t\,(S_{t+1}\,|R^i)$ are the expected future gold prices conditional upon $i = 0, 1$, no government sales or government sales, respectively, and p_t is the market's assessed probability that the government will sell gold. Flood and Garber (1980) examined the price level effects resulting from anticipated monetary reforms in hyperinflation-era Germany. In this case, the spot rate represents the price level, $E_t(S_{t+1}\,|R^i)$ are the expected future price levels conditional upon no reform and reform, alternatively, and p_t is the market's assessed probability that the reform will take place. Lewis (1991) evaluated the term structure of US interest rates following the 1979 change in Federal Reserve operating procedures to determine whether the market believed a shift in policy to lower interest rates was possible. In this case, S_t represents the interest rate, $E_t(S_{t+1}\,|R^1)$ is the expected future interest rates conditional upon on shift to lower rates, and p_t is the marker's assessed probability that this shift will take place. Bates (1991) used option prices to estimate the market's beliefs that the US stock market might crash before October 1987. In this case, S_t represents the stock price, $E_t(S_{t+1}\,|R^i)$ is the expected future stock prices conditional upon no crash or crash, respectively, and p_t is the market's assessed probability that the crash will occur. Bekaert et al. (2001) analysed international term structure returns using expectations of discrete shifts in short-term interest rate regimes. In this case, S_t is the excess return of long bonds over shortterm bonds, and R^i refer to different short-term interest rate regimes. Ang et al. (2007) examine the effects upon long-horizon initial public offering (IPO) returns based upon uncertainty about which performance regime determines a given initial listing. In this case, S_t refers to the abnormal returns and R^i dictate whether they follow under- or over-performance.

In general, when traders believe a future shift may occur in determinants of asset prices, expectations will have the form given in (2′), as the above examples demonstrate. Now suppose that no change in regime occurs in the sample. Define $(S_{t+1}\,|R^0)$ as observations drawn from the current regime process. Then, the forecast errors become:

$$\begin{aligned} u_{t+1} &= \left(S_{t+1}|R^0\right) - E_t S_{t+1} \\ &= \left[\left(S_{t+1}|R^0\right) - E_t\left(S_{t+1}|R^0\right)\right] \\ &\quad + p_t\left[E_t\left(S_{t+1}|R^0\right) - E_t\left(S_{t+1}|R^1\right)\right] \end{aligned} \tag{3′}$$

As long as the process does not change, the first term represents the forecast error conditioned on the current regime and therefore has mean zero. By contrast, the second term captures the effect of an expected switch to process R^1 that does not materialize in the sample. If the expected price conditioned on process R^0 is on average greater, say, than the price conditioned on regime R^1, the mean of the forecast errors within the sample will tend to be positive. Note that, for the Mexican peso example, the conditional expectations are simply constants where $E_t(S_{t+1}\,|R^i) = S^i$, for $i = 0, 1$, so that Eqs (2′) and (3′) are equivalent to (2′) and (3′) in this case. In general, however, the expectation conditional upon each regime varies over time as new information arrives to the market.

The example in (3′) illustrates the peso problem effects upon realized returns when no switches occur in the sample. Of course, the forecast error will include this event when the switch occurs. If the switches do not occur with sufficient frequency in the sample, however, forecast errors may continue to appear to be biased. Moreover, even with sufficient occurrences of these shifts, the forecast errors may be serially correlated since they weight the difference between the two expected processes, given by the second term on the right-hand side of Eq. (3′). When the probabilities or the differences in expectations under the two regimes are serially correlated, these components of the forecast errors are serially correlated as well. In this case, the difference between the spot rate and the forward rate as in (1) will be serially correlated even in the absence of risk premia. This explanation is consistent with the observation in Rogoff (1977) that Mexican peso futures prices before the devaluation did not follow a martingale as they should have by the efficient markets hypothesis.

The Peso Problem and Bayesian Learning

The simple intuition of the Mexican foreign exchange devaluation example casts the peso problem as a problem arising from anticipated *future* shifts in fundamentals. More generally, the peso problem phenomenon has also come to encompass the asset price implications due to uncertainty about *past* discrete changes. To see why the asset price behaviour is similar, consider a simple example. Suppose that market participants believe that the regime may have shifted in some past time period, $\tau < t$. Given priors about the probability of a change, they will then update their assessed probabilities of living in a new regime as new information arrives. If they learn through Bayesian inference, the forecast errors will depend upon expectations conditioned on each regime process and upon the updated probabilities of being in each regime.

The form of these forecast errors is isomorphic to Eq. (3′). To illustrate, suppose that in fact the process changed at time τ. In this case, the current regime R^0 is the new regime, and the alternative regime R^1 is the old regime. The probability p_t represents the market's assessed probability that no change took place. Over time, as the market learns the truth, the probability of no change goes to zero and the second component in the forecast error (3′) vanishes. Clearly, these forecast errors converge to mean-zero, white-noise levels even though they may appear biased during the learning process. Similar results hold when the market does not know the parameters of the new distribution but learns them over time. For example, Lewis (1989) relates the US dollar foreign exchange rate behaviour in the early 1980s to the market's uncertainty about whether a past shift to tighter monetary policy took place. Similarly, Timmermann (1993) shows how the learning can help explain the excess volatility in stock markets.

Despite the similarity of expectations based on learning about past discrete changes and on anticipating future discrete changes, their implications for forecast error behaviour in sufficiently large samples can be somewhat different. A once-and-for-all shift in the asset process with subsequent learning will induce forecast errors that are biased and serially correlated over the learning period. However, as the market learns, the probability of the old regime continuing will go to zero and the effect from the second term on the right-hand side of (3) will vanish. Thus, with sufficient observations, forecast errors following learning will behave according to the standard rational expectations assumptions; that is, they will be mean zero and serially uncorrelated. By contrast, with sufficient observations of the discrete shifts in processes, forecast errors arising from anticipated future discrete events will remain serially correlated in general but will be unbiased.

Empirical Approaches to the Peso Problem in Asset Prices

As this description makes clear, the peso problem is inherently a problem of identifying a low probability event in a given sample. Many researchers simply acknowledge that this small sample problem may be an issue in their results. Other researchers examine the potential for peso problems to explain anomalous asset price behaviour by using different approaches to identify the peso problem in sample.

These approaches can be divided into three main groups. The first group uses a calibrated asset pricing model to consider whether a peso problem explanation can explain a given empirical regularity. For example, Rietz (1988) uses this approach to consider whether the equity premium can be explained by rare adverse events. More recently, Barro (2006) examines the plausibility of this explanation using data over the 20th century.

The second group identifies the peso problem by using dates of known discrete changes in fundamentals to empirically back out expectations from asset prices. This group of studies focuses upon easily observable shifts in fundamentals. Examples include exchange rate realignments (Bertola and Svensson 1993; Campa and Chiang 1996; Campa et al. 2002; Mundaca 2004) and announced shifts in monetary policy targeting (Lewis 1991; Hallwood et al. 2000).

The third group analyses the peso problem by directly estimating regimeswitching models of fundamentals to explain anomalous behaviour in their asset prices. This approach has the advantage that the fundamentals process can be estimated from the available data and does not require the researcher to take a stand on the timing of the events. As a result, the analysis can be conducted in a wide range of applications where the dating of events is not known a priori. Many different asset prices have been studied using this approach, including floating spot exchange rates (Engel and Hamilton 1990; Kaminsky 1993), the equity premium (Cecchetti et al. 1993), the real interest rate (Evans and Lewis 1995a), the foreign exchange risk premium (Evans and Lewis 1995b), the term premium (Bekaert et al. 2001), and IPO abnormal returns (Ang et al. 2006).

Summary

In summary, as long as agents anticipate occasional discrete changes in the process of economic variables that affect asset prices, and these changes occur infrequently, asset prices contain the potential for the peso problem. If so, then forecast errors will be serially correlated. Furthermore, unless the sample contains many observations of the discrete shifts, forecast errors will appear biased when observed ex *post* even though traders may have rational expectations. Despite this problem, empirical financial studies frequently measure the risk premium as the predictable component of the realized spot rate less the forward rate, described in (1). Therefore, if the 'peso problem' is present in the sample, researchers may incorrectly attribute asset price behaviour to anomalies rather than to the market's rational forecasts of discrete events.

See Also

- ▶ Bayesian Statistics
- ▶ Currency Crises
- ▶ Financial Market Anomalies
- ▶ Finite Sample Econometrics
- ▶ International Finance
- ▶ Rational Expectations
- ▶ Regime Switching Models

Bibliography

Ang, A., L. Gu, and Y. Hochberg. 2007. Is IPO underperformance a peso problem? *Journal of Financial and Quantitative Analysis* 42: 565–594.

Barro, R. 2006. Rare disasters and asset markets in the 20th century. *Quarterly Journal of Economics* 121: 823–866.

Bates, D.S. 1991. The Crash of 87: was it expected? The evidence from options markets. *Journal of Finance* 46: 1009–1044.

Bekaert, G., R.J. Hodrick, and D.A. Marshall. 2001. Peso problem explanations for term structure anomalies. *Journal of Monetary Economics* 48: 241–270.

Bertola, G., and L.E.O. Svensson. 1993. Stochastic devaluation risk and the empirical fit of target-zone models. *Review of Economic Studies* 60: 689–712.

Campa, J.M., and P.H.K. Chang. 1996. Arbitrage-based tests of target-zone credibility: Evidence from ERM cross-rate options. *American Economic Review* 86: 726–740.

Campa, J.M., P.H.K. Chang, and J.F. Refalo. 2002. An options-based analysis of emerging market exchange rate expectations: Brazil's real plan, 1994–1999. *Journal of Development Economics* 69: 227–253.

Cecchetti, S.G., P. Lam, and N.C. Mark. 1993. The equity premium and the risk-free rate: Matching the moments. *Journal of Monetary Economics* 31: 21–45.

Engel, C., and J.D. Hamilton. 1990. Long swings in the dollar: Are they in the data and do markets know it? *American Economic Review* 80: 689–713.

Evans, M.D.D., and K.K. Lewis. 1995a. Do expected shifts in inflation affect estimates of the long-run Fisher relation? *Journal of Finance* 50: 225–253.

Evans, M.D.D., and K.K. Lewis. 1995b. Do long-term swings in the dollar affect estimates of the risk premia? *Review of Financial Studies* 8: 709–742.

Flood, R., and P. Garber. 1980. An economic theory of monetary reform. *Journal of Political Economy* 88: 24–58.

Hallwood, C.P., R. MacDonald, and I.W. Marsh. 2000. Realignment expectations and the US dollar, 1890–1897: was there a 'Peso Problem'? *Journal of Monetary Economics* 46: 605–620.

Kaminsky, G. 1993. Is there a peso problem? Evidence from the dollar/pound exchange rate: 1976–1987. *American Economic Review* 83: 450–472.

Krasker, W.S. 1980. The peso problem in testing the efficiency of forward exchange markets. *Journal of Monetary Economics* 6: 269–276.

Lewis, K.K. 1989. Changing beliefs and systematic forecast errors. *American Economic Review* 79: 621–636.

Lewis, K.K. 1991. Was there a peso problem in the U.S. interest rate term structure: 1979–1982. *International Economic Review* 32: 159–173.

Lizondo, J.S. 1983. The efficiency of the Mexican peso market. *Journal of International Economics* 14: 69–84.

Mundaca, G. 2004. Modeling probabilities of devaluations. *Economica* 71: 13–37.

Rietz, T.A. 1988. The equity risk premium: A solution. *Journal of Monetary Economics* 22: 117–131.

Rogoff, K. 1977. *Rational expectations in the foreign exchange market revisited*. Massachusetts Institute of Technology. Unpublished manuscript.

Rogoff, K. 1980. Tests of the martingale model for foreign exchange futures markets. In *Essays on expectations and exchange rate volatility*. Ph.D. dissertation, Massachusetts Institute of Technology.

Salant, S., and D. Henderson. 1978. Market anticipations of government policies and the price of gold. *Journal of Political Economy* 86: 627–648.

Timmermann, A.G. 1993. How learning in financial markets generates excess volatility and predictability in stock prices. *Quarterly Journal of Economics* 108: 1135–1145.

Peter, Hans (1898–1959)

H. C. Recktenwald

After initial studies at Tübingen in theology, mathematics and philosophy, and later on in economics, Peter became Lectúrer (*Privatdozent*) in economics and statistics at the University of Tübingen. The Nazi regime denied him a chair because his thorough but critical analysis of Marx's and Ricardo's theories was considered to show dependence on Jewish thinking. Peter courageously defended his intellectual and ethical position as a scientist in an open letter to the *Finanz-Archiv*, a valuable document of this dark period of German history. It was not until 1947 that he obtained a full professorship in Tübingen, where he died in 1959.

Peter had a sharp and independent mind. His scientific interests and deep knowledge ranged from abstract philosophy, logic and ethics to economic analysis and social policy. He applied relatively complicated mathematical models to explain the circular flow of the economy and was one of the first economists in Germany to use econometric methods and game theory. He also made an essential contribution to modern growth theory. He postulated the application of a variety of methods (a *Methodenpluralismus*). It seems justified to classify Peter as a liberal socialist, whatever this contradictory notion may mean.

Selected Works

1933–7. *Grundprobleme der theoretischen Nationalökonomie*, 3 vols. Bonn: Schroeder.
1950. *Einführung in die politische Ökonomie*. Stuttgart/Cologne: Kohlhammer.
1954. *Mathematische Strukturlehre des Wirtschaftskreislaufes*. Göttingen: Schwarz.

References

Recktenwald, H.C. 1985. *Ordnungstheorie und ökonomische Wissenschaft*. Erlangen: Universitätsbund.

Petty, William (1623–1687)

Alessandro Roncaglia

Keywords
Cantillon, R.; Classical value theory; Demography; Economic statistics; Expenditure tax; Fiscal pressure; Inflation tax; Just price; Labour theory of value; Market price; Marx, K. H.; Maximum wages; Money; Natural price; Optimal taxation; Petty, W.; Political anatomy; Political arithmetic; Quesnay, F.; Rent; Subsistence wage; Surplus; Velocity of circulation; Wealth

JEL Classifications
B31

Sir William Petty was born on 26 May 1623 in the village of Romsey, Hampshire, and died on 26 December 1687 in London. His life was hectic: son of a clothier, he was a cabin boy on a merchant ship at 13, admitted to the Jesuit college in Caen (France) at 14; after serving in the Royal Navy he sought refuge in the Netherlands (1643) and Paris (1645), where he studied medicine and (with Hobbes) anatomy. He returned to Romsey in 1646 to revive his father's business; became a doctor of medicine in Oxford University in 1648, and, after an impressive academic career, Professor of Anatomy in 1650, but moved immediately – in 1651 – to the Chair of Music at Gresham College, London. He was also appointed chief medical officer to the English army in Ireland in 1651, and was responsible in 1655–8 for the topographical survey of Irish lands destined for English soldiers, from which he himself emerged with a large landed estate. From then until his death, he was engaged in the management of his estate and in endless litigation over titles of property and taxes, constantly travelling between England and Ireland. Petty also managed to participate, in 1660–2, in the founding of the Royal

Society (in full: the Royal Society for the Improving of Natural Knowledge) and in furthering its activities. He married Elizabeth Waller in 1667, and had five children by her and at least one illegitimate child.

Only a small part of Petty's written work was published under his own name during his lifetime. The main essays concerned with economic issues were published after his death, soon after the Glorious Revolution of 1688 made the political climate more favourable to the reception of Petty's ideas. The *Verbum Sapienti* and the *Political Arithmetick* were published in 1690, The *Political Anatomy of Ireland* in 1691, and the *Quantulumcumque concerning Money* in 1695, though they were written respectively in 1664, 1676, 1672 and 1682. Among the writings published during Petty's lifetime, the *Natural and Political Observations upon the Bills of Mortality* appeared in 1662 under the name of John Graunt, a good friend, although it seems certain that Petty authored at least part of it.

This work is generally considered as marking the birth of the science of demography. A collection of Petty's economic writings, containing some unpublished material, appeared in 1899 as *The Economic Writings of Sir William Petty*, edited by Charles Hull. In 1927 and 1928 other unpublished material appeared (*The Petty Papers*, in two volumes; and *The Petty–Southwell Correspondence*), edited by the sixth Marquis of Lansdowne, a descendant of Petty. Unpublished material (known as 'the Bowood Papers') is still extant at the Bodleian Library in Oxford. An important item – *A Dialogue on Political Arithmetic* – edited by S. Matsukawa was published in 1977 (Matsukawa 1977). (For Petty's complete bibliography, see Keynes 1971; for a bibliography on Petty, see Roncaglia 1985.)

Petty's contribution to the origins of classical political economy is threefold, involving method, conceptual framework, and analysis. These aspects are interconnected and often implicit in Petty's writings, which specifically refer to policy issues of his time. We will consider the three aspects separately for the sake of clarity, summarizing Petty's ideas on each of them from his several writings.

Petty refers to his method as 'political arithmetick', which comprises the following principles:

> To express my self in Terms of Number, Weight or Measure; to use only Arguments of Sense, and to consider only such Causes, as have visible Foundations in Nature; leaving those that depend upon the mutable Minds, Opinions, Appetites and Passions of particular Men, to the Consideration of others. (Petty 1899, p. 244)

This method recalls Hobbes's *logica sive computatio*, and Bacon's inductive method (to which Petty explicitly refers). It points to a rejection of the then prevailing qualitative approach to science, based on the description of the quality of the sensations associated with physical objects and human events, in favour of the newly rising quantitative–objectivistic approach. The physical sciences were experiencing this shift during the 17th century; the foundation of the Royal Society marked a decisive step in the transition from the old to the new methodology.

In this respect, the commonly held idea that Petty's 'political arithmetick' simply marks the origin of modern economic statistics should be rejected. Petty aims at something more than *recording and describing reality* 'in terms of number, weight or measure'. He aims at *expressing reality* in such terms, since this allows him to identify 'such causes, as have visible foundations in nature', that is, to identify the laws intrinsic in reality. Petty thus adopts a point of view which was embedded in the new quantitative approach to science. As Galileo expresses it: 'This great book which is open in front of our eyes – I mean the Universe – . . . is written in mathematical characters' (Galilei 1623, p. 232). According to this point of view reality *contains* natural laws, so that the task of the scientist is to *discover* these natural laws lying beneath the surface of the apparently erratic phenomena experienced by our senses. Petty himself recognizes that as a description of reality political arithmetick is necessarily imperfect; however, his aim is to locate reality's *inner structure*, not the descriptive details.

Petty's methodological contribution to the development of political economy consists precisely in this: he brings the new quantitative method into the political science dealing with the nature and causes of social wealth. As already stressed, this method means more than an intention to *measure* social phenomena: it means a systematic search for the main characteristics of human societies – a fact well expressed by Petty's other favourite term for the object of his enquiries, 'political anatomy'.

In the Preface to *The Political Anatomy of Ireland* (1691) Petty recalls Francis Bacon's parallel between the 'body natural' and the 'body politick'. (This parallel has a long tradition indeed, going back to Menenio Agrippa's apology in Ancient Rome; but in Petty's writings it loses its old moral connotation, no longer suggesting the need for diverse social groups to cooperate.) Petty then goes on: 'as Anatomy is the best foundation of the one [the body natural], so also of the other [the body politick]', and points to the need of 'knowing the Symmetry, Fabrick, and Proportion' of the 'Body Politick' (Petty, 1899, p. 129).

There is a clear parallel between the triad 'Symmetry, Fabrick and Proportion' of Political Anatomy, and the triad) 'Number, Weight or Measure' of Political Arithmetick. The new science, of which Petty claims to be the founder, is thus characterized both by its quantitative nature and by its objectivistic approach. Also, Petty's reference to the Political Body points to the 'systemic nature' of the new approach (or, in other terms, to the 'holistic nature' which from Petty onwards characterizes classical political economy).

The vision of society as a political body, comparable to the human body, was probably influenced by Petty's medical career, which earned him the Oxford chair in Anatomy (another illustrious example of a doctor–economist is provided by the founder of the Physiocratic school, François Quesnay). Thus the human body–political body comparison constitutes the background to the specification of the conceptual framework of the new science, to which Petty makes an important contribution.

Petty's notion of money, for instance, is specified through a human-body metaphor:

> Money is but the Fat of the Body-politick, whereof too much doth as often hinder its Agility, as too little makes it sick . . . As Fat lubricates the motion of the Muscles, feeds in want of Victuals, fills up uneven Cavities, and beautifies the Body, so doth Money in the State quicken its Action, feeds from abroad in the time of Dearth at Home; evens accounts by reason of its divisibility. (Petty 1899, p. 112)

This metaphor shows that Petty perceived the three functions of money: unit of measure, means of exchange, and store of value. It also shows that Petty did not consider money as constituting the wealth of nations. In fact, Petty's notion of wealth is well expressed through another body-politick metaphor (in all likelihood influenced by William Harvey's then recent discovery of the circulation of the blood: 'the blood and nutritive juices of the Body Politick' are 'the product of Husbandry and Manufacture' (Petty 1899, p. 28).

In relation to money, we can also recall that Petty clearly perceived the notion of the velocity of circulation (which is measured through reference to the customary payment intervals for taxes, rents, wages; and which is used for estimating the optimal quantity of money required to finance a given volume of income and trade).

Furthermore, Petty stresses that banks, in creating paper money, allow society to save on the cost of acquiring the precious metals necessary for ensuring the required monetary circulation.

The idea of the 'body-politick' is also relevant for Petty's analysis of the fiscal system, which mainly concerns its impact on the economic development of society.

Petty confronts his ideas on the optimal conditions for a system of taxation, considered as a coherent whole, with the chaotic situation then prevailing, and spells out the preconditions for modern fiscal institutions. He also introduces the notion now known as fiscal pressure, frequently referring in his works to the ratio between the amount of taxation and the level of national income (or of national expenditure: in fact, Petty favoured an expenditure tax over an income tax system). Interestingly, but not uncommonly, the devaluation of the currency (that is, inflation) is considered as a particular kind of tax.

The idea of the 'body-politick' implies a connection between the concept of the economic system and the concept of the nation-state. Machiavelli's work most probably influenced Petty in this respect, as well as in the shift from the moral judgement of human actions to the objective analysis of social events. Machiavelli and Petty also share common limits in their notions of the nation-state and the economic system, since they seem not to perceive the productive interrelationships connecting city and countryside, industry and agriculture: productive interrelationships on which Richard Cantillon was to focus attention, and which would constitute the main analytical contribution of Quesnay's *Tableau économique* in the 18th century.

The notion of the surplus is generally regarded as one of Petty's most relevant contributions. Petty expresses the surplus in physical terms, as the amount of product (corn) exceeding the required means of production, and identifies it with rent. In this way Petty avoids the problem of the determination of the profit rate, which in turn involves the problem of relative prices, since relative prices are required for evaluating both capital advances and the net product. (Such problems were to be taken up later by classical economists like Ricardo and Marx, and then, more recently, by Piero Sraffa.) Interestingly, Petty also expresses the surplus in terms of the number of unemployed persons who can be maintained by a group of labourers who are producing the strict necessaries for both groups, workers and non-workers alike: shades of Marx's surplus labour notion? Like the production of services and luxury goods, unemployment thus appears as a particular way of utilizing the surplus. Wages are not included in the surplus, since they correspond to the necessary subsistence of the workers (and Petty, who considers the workers as nothing else but a produced means of production, considered the subsistence wage not as the result of some automatic mechanism, but as an objective to be reached through laws regulating maximum wages).

Petty's strictly analytical contributions to the origins of classical political economy are more limited than his methodological and conceptual contributions, but are nonetheless relevant.

Petty was credited, by Marx and others, with a labour-embodied theory of value. However, the passages usually quoted to support this interpretation are in fact simplifications of a more complex (and less useful) labour-cum-land theory, based on the idea that the price of each commodity depends on the quantities of the various means of production required to obtain it. In particular, the absence of any consideration of profits and the profit rate suggests that Petty's theory of prices must be considered as very primitive. However, it provided a starting point for subsequent developments. Richard Cantillon's posthumously published *Essay* (1755), for example, dwells on the problem of the 'par' between labour and land, and this is derived from Petty's attempt to find a way of expressing one of the two 'originary' means of production in terms of other, in order to obtain a single magnitude expressing the difficulty of production of any commodity.

But what is especially relevant for all subsequent analyses of price is Petty's distinction between 'actual' and 'natural' prices or, in other terms, between exchange relationships actually taking place, and theoretical prices which express the most relevant factors influencing current prices. Petty clearly identifies (in the *Dialogue of Diamonds*, first published in 1899) the factual preconditions – the existence of a regular market, namely, of repeated acts of exchange following regular patterns – necessary for the notion of 'natural price' to be meaningful. This is an *objective* notion of natural price, distinct from the notion of the 'just price', the determination of which, as a moral rule of behaviour for sellers and buyers, was one of the main purposes of the writers dealing with economic issues for centuries before Petty's time (for example, Pufendorf). Thus, once again, Petty's contribution to the development of classical political economy relates more to concepts and method than to specific analytic propositions. Nevertheless, it is difficult to overestimate his contribution to classical value theory, for which the notion of natural price (as well as the surplus) represents a necessary prerequisite.

Petty's importance for 17th-century culture is undeniable. His search for an 'objective' science contributes to the paradigm shift that was taking place at the time. In this regard his part in the creation of the Royal Society went hand in hand with his development of the new science of 'political arithmetick'. The 'human body–political body' comparison provides a much-needed 'systemic' background to the emerging objective analysis of economic events. On both levels ('political arithmetick' and 'political anatomy'), his influence on subsequent developments was decisive: his immediate followers (for example, Gregory King and Charles Davenant) definitively established the sciences of demography and economic statistics, while Petty's conceptual framework, adopted by Cantillon, exerted a decisive influence on the development of Quesnay's economic thinking. In this way (as well as through other less direct channels) Petty influenced both Smith and Ricardo, even if they do not refer directly to his writings. Petty's relevance for the development of classical political economy is emphasized by Karl Marx, who considers Petty to be the 'founder of Classical political economy'. Later economists limit reference to some specific aspect of Petty's ideas: for instance Keynes (1936, pp. 359, 362) quotes with approbation his ideas on the use of public works as a tool of employment policy; and Luigi Einaudi (1941) refers with enthusiasm to Petty's preference for expenditure taxes. However, these aspects, while testifying to Petty's brilliant intelligence, should not obscure what are in fact his main contributions to economic science: the emphasis on the 'objective' method, and the establishment of certain key concepts which later became so basic to economic science as to be unconsciously but consistently accepted as part of our scientific background.

See Also

- ► Davenant, Charles (1656–1714)
- ► Political Arithmetic

Selected Works

1690a. *Verbum Sapienti*. Reprinted in Petty (1899).

1690b. *Political Arithmetick*. Reprinted in Petty (1899).

1691. *The political anatomy of Ireland*. Reprinted in Petty (1899).

1695. *Quantulumcumque concerning Money*. Reprinted in Petty (1899).

1899. *Economic writings*. 2 vols, ed. C. Hull. Reprinted, New York: A.M. Kelley, 1964.

1927. *Papers*. 2 vols, ed. H. Lansdowne. London: Constable.

1928. *The Petty–Southwell Correspondence, 1676–1687,* ed. H. Lansdowne.

London: Constable.

Bibliography

Cantillon, R. 1755. *Essai sur la nature du commerce en général*. Ed. H. Higgs, London: Macmillan, 1931.

Einaudi, L. 1941. *Saggi sul risparmio e l'imposta*. Turin: Einaudi.

Fitzmaurice, E. 1895. *The life of Sir William Petty*. London: Murray.

Galilei, G. 1623. *Il Saggiatore*. Reprinted in *Le opere*, vol. 6, ed. A. Favaro. Florence, 1890–1909.

Graunt, J. 1662. *Natural and political observations upon the bills of mortality*. Reprinted in Petty (1899).

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

Keynes, G. 1971. *A bibliography of Sir William Petty, F.R.S. and of observations on the bills of mortality by John Graunt, F.R.S.* Oxford: Clarendon Press.

Matsukawa, S. 1977. Sir William Petty: An unpublished manuscript. *Hitotsubashi Journal of Economics* 17 (2): 33–50.

Roncaglia, A. 1985. *Petty: The origins of political economy*. New York: Sharpe.

Pharmaceutical Industry

Sara Fisher Ellison

Keywords

Agency problems; Asymmetric information; Entry; Industrial organization; Innovation; Patents; Pharmaceutical industry; Rregulation; Research

JEL Classifications

L65

The pharmaceutical industry comprises firms which manufacture medicines, including vaccines. Many of these firms also perform some or all of the following functions: conducting basic scientific research to identify (patentable) chemical compounds with medicinal properties, developing those chemical compounds into safe, effective, and commercially viable medicines, gaining government approval to sell those medicines, and marketing those medicines to potential consumers and prescribers. This industry has been widely studied by those interested in the analysis of health care systems. However, I focus here on industrial organization research which seeks to explain general economic phenomena by using the pharmaceutical industry as a setting.

Certain salient features of the pharmaceutical industry have made it a popular focus of research in industrial organization. First, asymmetric information and agency problems are present. Second, innovation plays a central role. Third, entry of new products is common. Fourth, data are unusually available. Finally, the industry is regulated along many dimensions.

Economists have used the pharmaceutical industry to study how asymmetric information and agency problems can affect demand for products in a differentiated product setting. Properties of medicines are not always easily verified or understood by consumers. Furthermore, some medicines are available only through a physician's prescription, and consumers may not be the ones making purchase decisions or paying for the medicine once the decision is made. Hellerstein (1998) and Stern and Trajtenberg (1998) study the role of the prescribing physician in the type of medicine dispensed. Ellison et al. (1997) measure price sensitivity of various agents involved in prescribing and dispensing medicines. Berndt et al. (2003) study the impact of incomplete product information on the diffusion of medicines after initial release.

Researchers have used the industry to study determinants of innovation, including incentives provided to firms by various patent systems, incentives faced by researchers within a firm, the size of the firm's research effort, the diversity of its research portfolio, and the geographic proximity of other research centres. Work in this area includes Henderson and Cockburn (1996) and Azoulay (2004).

Most pharmaceutical products are initially patent-protected because they are based on the discovery or synthesis of some new chemical compound. When patents expire, then, the potential exists for entry by chemically identical products, or 'generics'. The large number of similar markets with observable dates of potential entry has proven a boon to researchers studying entry. Caves et al. (1991) and Scott Morton (1999) identify factors important in generic manufacturers' decisions to enter a market, and Ellison and Ellison (2000) look for empirical evidence of strategic entry deterrence by incumbent producers. Pervasive entry has also made the industry a natural setting for critiquing how government price indices handle product introductions. Griliches and Cockburn (1995) and Berndt et al. (1993) influenced the Boskin Commission report (Boskin et al., 1996), which suggested alternative ways of computing those indices.

Study of vertical relationships is often hampered by the proprietary nature of the transactions between firms. But pharmaceutical wholesale transactions data are often available, enabling studies such as Ellison and Snyder (2001), which tests various theories of buyer size effects.

Past regulation has shaped the industry, and significant effort is expended by the industry to shape future regulation in turn. Ellison and Mullin (2001) demonstrate the effect on the industry of proposed regulatory reform in the early 1990s, while Ellison and Wolfram (2000) provide evidence of actions the industry took to forestall reform. Also, Scott Morton (1997) studies the distortionary effect of government procurement regulations on firms' pricing decisions. Much of the research on the pharmaceutical industry has focused on the United States, but interesting questions involving international comparisons of regulatory regimes have been addressed by Danzon and Chao (2000), focusing mainly on price differences, and Kyle (2005), focusing on firms' entry, or 'launch' decisions.

Bibliography

Azoulay, P. 2004. Capturing knowledge across and within firm boundaries: Evidence from clinical development. *American Economic Review* 94: 1591–1612.

Berndt, E., Z. Griliches, and J. Rosett. 1993. Auditing the producer price index: Micro evidence from prescription pharmaceutical preparations. *Journal of Business and Economic Statistics* 2: 251–264.

Berndt, E., R. Pindyck, and P. Azoulay. 2003. Consumption externalities and diffusion in pharmaceutical markets: Antiulcer drugs. *Journal of Industrial Economics* 51: 243–270.

Boskin, M., E. Dulberger, R. Gordon, Z. Griliches, and D. Jorgenson. 1996. *Toward a more accurate measure of the cost of living.* Final report to the Senate Finance Committee from the Advisory Commission to Study the Consumer Price Index. Washington, DC: Senate Finance Committee.

Caves, R., M. Whinston, and M. Hurwitz. 1991. Patent expiration, entry and competition in the US pharmaceutical industry: An exploratory analysis. *Brookings Papers on Economic Activity* 1991: 1–48.

Danzon, P., and L. Chao. 2000. Cross-national price differences for pharmaceuticals: How large and why? *Journal of Health Economics* 19: 159–195.

Ellison, G., and S. Ellison. 2000. *Strategic entry deterrence and the behavior of pharmaceutical incumbents prior to patent expiration*. Mimeo. Cambridge, MA: MIT.

Ellison, S., and W. Mullin. 2001. Gradual incorporation of information: Pharmaceutical stocks and the evolution of President Clinton's health care reform. *Journal of Law and Economics* 44: 89–129.

Ellison, S., and C. Snyder. 2001. *Countervailing power in wholesale pharmaceuticals*, Working Paper 01-27. Cambridge, MA: MIT.

Ellison, S., and C. Wolfram. 2000. *Pharmaceutical prices and political activity*, Working Paper 8482. Cambridge, MA: NBER.

Ellison, S., I. Cockburn, Z. Griliches, and J. Hausman. 1997. Characteristics of demand for pharmaceutical products: An examination of four cephalosporins. *RAND Journal of Economics* 28: 426–446.

Griliches, Z., and I. Cockburn. 1995. Generics and new goods in pharmaceutical price indexes. *American Economic Review* 84: 1213–1232.

Hellerstein, J. 1998. Importance of the physician in the generic versus trade-name prescription decision. *RAND Journal of Economics* 29: 108–136.

Henderson, R., and I. Cockburn. 1996. Scale, scope and spillovers: Determinants of research productivity in the pharmaceutical industry. *RAND Journal of Economics* 27: 32–59.

Kyle, M. 2005. *Pharmaceutical price controls and entry strategies*. Mimeo, Duke University.

Scott Morton, F. 1997. The strategic response by pharmaceutical firms to the Medicaid most-favored-customer rules. *RAND Journal of Economics* 28: 269–290.

Scott Morton, F. 1999. Entry decisions in the generic drug industry. *RAND Journal of Economics* 30: 421–440.

Stern, S., and M. Trajtenberg. 1998. *Empirical implications of physician authority in pharmaceutical decision-making*, Working Paper 6851. Cambridge, MA: NBER.

Phelps Brown, (Ernest) Henry (1906–1994)

Guy Routh

Keywords

Inequality; Phelps Brown, E. H.; Wage rates

JEL Classifications

B31

Born in Calne, Wiltshire, on 10 February 1906, Phelps Brown was educated at Taunton School and then at Wadham College, Oxford, where he was a Scholar and gained First Class Honours in Modern History (1927) and in Philosophy, Politics and Economics (1929). He was a Fellow of New College from 1930 to 1947. In 1936 he published *The Framework of the Pricing System*, an orthodox exposition of marginal theory notable for its clarity.

After distinguished war service with the Royal Artillery (which provided material for *The Balloon*, a novel published in 1953), he became the first Professor of the Economics of Labour at the University of London, teaching at the London School of Economics from 1947 until 1968, when he retired as Emeritus Professor. His lecture courses, 'Applied Economics' and 'The Economics of Labour', were well attended and valued for their incisiveness and lucidity. *A Course in Applied Economics* was published in 1951, frequently reprinted and issued in a second edition, with J. Wiseman, in 1964. The reader was invited to apply economic analysis to practical problems 'seen in the many-sidedness that calls for more insights than those of the economist alone'. *The Economics of Labor* appeared in 1962 as the first of the Yale University Studies in Comparative Economics. At the LSE, Phelps Brown carried out a series of studies in the tradition of the great British sociologist-statisticians. These were republished in Henry Phelps Brown and Sheila V. Hopkin, *A Perspective of Wages and Prices* (1981). They are characterized by a scrupulous assembly of data from various countries and, in the case of building wages and the price of consumables, extended over seven centuries. A remarkable stability was found in building wage rates, with no sustained change in 500 out of 690 years, and in differentials between craftsmen and labourers, with a failure by supply and demand to overcome 'the inertia of convention' (p. 8). An ability to combine history, sociology and statistics to illuminate economics is demonstrated in *The Growth of British Industrial* Relations: *A Study from the Standpoint of 1906–14* (1959), in *A Century of Pay* (1968) and in *The Inequality of Pay* (1977). In this last, a mass of data from many countries is marshalled and analysed to assess the relative significance of market and sociological factors in determining inequalities of pay. The conclusion is that these differences are 'better explained by the play of market forces than by that of custom, convention, status, or power' (p. 325). In *The Origins of Trade Union Power* (1983), however, the emphasis is on socio-psychological forces: 'because attitudes govern responses, they are among the basic determinants of the course of history . . . . At the last we are left with the paradox of historical understanding, that we can trace past happenings to their causes without thereby gaining the power to predict' (pp. 300 and 302). Phelps Brown pursued his work on inequality in the wide ranging *Egalitarianism and the Generation of Inequality* (1988).

Phelps Brown served on a number of public bodies: as one of the 'Three Wise Men' (the Council on Prices, Productivity and Income) in 1959; on the National Economic Development Council, 1962; on the OECD Working Party on Wages and Labour Mobility, 1963–4; and on the Royal Commission on the Distribution of Income and Wealth, 1974–8. He was awarded a knighthood in 1976, and became a Fellow of the British Academy in 1960. He was President of the Royal Economic Society, 1970–2, and in his presidential address (published in the *Economic Journal*, 1972) presented his credo on the nature and methods of economics, joining other critics who had independently arrived at similar conclusions: training in advanced economics might be actively unhelpful to those concerned with the application of policy, for 'it is impaired from the first by being built upon assumptions about human behaviour that are plucked from the air' (p. 3). His remedies were the removal of the traditional boundary between economics and the other social sciences; a clinical commitment to diagnose and prescribe for particular economic ailments, beginning with practice and working back to theory; the study of history as an essential part of economic training; more observation of actual behaviour, ingenuity in devising methods, accumulating facts, seeking connections and significant detail (p. 9). This analysis was further developed in 'The Radical Reflections of an Applied Economist' (1980), reinforcing and extending the arguments of 1972.

A detailed obituary is Hancock and Isaac (1998); see as well his own 'Autobiographical Notes' (1996).

Selected Works

1936. *The framework of the pricing system*. London: Chapman & Hall.

1951. *A course in applied economics*. London: Sir Isaac Pitman & Sons.

1959. *The growth of British industrial relations: A study from the standpoint of 1906–14*. London: Macmillan.

1962. *The economics of labor*. New Haven/London: Yale University Press.

1964. (With J. Wiseman.) *A course in applied economics*, 2nd ed. London: Sir Isaac Pitman & Sons.

1968. (With M.H. Browne.) *A century of pay: The course of pay and production in France, Germany, Sweden, the United Kingdom, and the United States of America, 1860–1960*. London: Macmillan.

1972. The underdevelopment of economics. *Economic Journal* 82: 1–10.

1977. *The inequality of pay*. Oxford: Oxford University Press.

1980. The radical reflections of an applied economist. *Banca Nazionale del Lavoro Quarterly Review* 132: 3–14.

1981. (With S.V. Hopkins.) *A perspective of wages and prices*. London: Methuen.

1983. *The origins of trade union power*. Oxford: Clarendon Press.

1988. *Egalitarianism and the generation of inequality*. Oxford: Clarendon Press.

1996. Autobiographical notes. *Review of Political Economy* 8: 129–139.

Bibliography

Hancock, K., and J.E. Isaac. 1998. Henry Phelps Brown 1906–1994. *Economic Journal* 108: 757–778.

Phelps, Edmund (Born 1933)

Gylfi Zoega

Abstract

Edmund Phelps is a Nobel Prize winner in economics who has contributed to our understanding of the supply side of the macroeconomy. He showed that there is no stable trade-off between inflation and unemployment. He derived the socially optimal level of saving and the socially optimal level of research into new technologies and showed how technological progress depended on the size of the population and its level of education. In recent years, Phelps has developed models of the equilibrium unemployment rate, what he calls structural unemployment, that can explain the long swings of unemployment as well as differences across countries.

Keywords

Asymmetric information; Demand shocks; Dynamic inefficiency; Economic growth in the very long run; Education; Efficiency wages; Endogenous growth theory; Expectations; Golden rule; Hyperbolic preferences; Hysteresis; Imperfect information; Inflation expectations; Intertemporal investment; Learning; Lucas, R; Matching; Microfoundations; Monetary shocks; Natural rate of unemployment; Natural rate of unemployment; New classical economics; New Keynesian macroeconomics; Path dependence; Phelps, E.S; Phillips curve; Productivity growth; Rational expectations; Stabilization policy; Statistical discrimination; Structural unemployment; Technical change; Unemployment–inflation trade-off

JEL Classifications

B31

Edmund S. Phelps was born in Evanston, Illinois, on 26 July 1933 and grew up in Hasting-on-Hudson, New York. He attended Amherst College as an undergraduate, where he took a second-year economics course, at his father’s suggestion, which sparked an interest in economics. After receiving his BA degree in 1955, Phelps started graduate studies at Yale, where he was influenced by, among others, James Tobin, William Fellner, Henry Wallich and Thomas Schelling. He received his Ph.D. from Yale in 1959. After a short spell at the Rand Corporation, Phelps accepted a research post at the Cowles Foundation in 1960. In 1966 he left Yale for the University of Pennsylvania where he stayed until 1969. There followed a year visiting Stanford University and then in 1971 a move to Columbia University, where he was later made McVickar Professor of Political Economy. At Columbia he met his wife, Viviana Montdor Phelps.

Phelps was elected to the National Academy of Sciences (USA) in 1981 and was made a Distinguished Fellow of the American Economic Association in 2000. He is also a former vice-president of the Association, a fellow of the Econometric Society, the American Academy of Arts and Sciences and the New York Academy of Sciences. A Festschrift conference in his honour was held at Columbia University in October 2001 and the volume published by Princeton University Press in 2003 (Aghion et al. 2003). Phelps received the Nobel Prize in economics in 2006 for his work on intertemporal trade-offs in macroeconomics.

The Economics of Phelps – A Brief Outline

During a telephone conversation with journalists in Stockholm, after being told that he had been awarded the Nobel Prize in economics, Phelps described his contribution as that of introducing people to macroeconomics. This was indeed the tenet of the 'Phelps volume' of 1970, which contained a selection of path-breaking papers, all providing microeconomic foundations for macroeconomics (Phelps et al. 1970c). By convincing others to follow suit in explaining macroeconomic relationships with models that describe the behaviour of firms, workers and consumers, Phelps began to transform macroeconomics. Moreover, he also contributed to bringing economic theory closer in line with 20th-century economic life by emphasizing imperfect information and imperfect knowledge with its accompanied market failures into macroeconomics. This heralded another transformation of the field.

Since the publication of his well-known paper on the golden rule of accumulation (Phelps 1961), Phelps has introduced new ways of thinking about such diverse issues as the effect of monetary policy on output and employment; equilibrium unemployment and efficiency wages; the sources of economic growth in the long run; imperfect competition; discrimination in the workplace; and optimal inflation targeting. His work can be divided chronologically into four distinctive phases. In the early to mid-1960s he wrote extensively on growth theory and produced the golden rules of growth and models of technological progress that were genuine precursors to what we now call endogenous growth theory. In the late 1960s and early 1970s his attention turned to the unemployment–inflation trade-off. Phelps showed how an increase in the supply of money would make firms raise output in the short run while in the long run only wages and prices were affected. The rejection of the notion of a stable Phillips curve – providing policymakers with a menu of unemployment/inflation pairs – was one of the most significant achievements in the history of macroeconomic thought, which changed the practice of monetary policy profoundly. It also opened the avenue to research on the optimal design of monetary policy, which essentially became an intertemporal optimization problem (see Phelps 1967, 1972b), as well as towards studying the determinants of the steady-state equilibrium unemployment rate, which Friedman dubbed the natural rate of unemployment (Friedman 1968). A third phase in Phelps's research consisted of a reaction to the rational expectations revolution and its challenge to the effectiveness of monetary policy. In the 1970s Phelps and colleagues – mainly at Columbia University – constructed models with rational expectations but also having wage and price contracts, wages and prices set for longer periods than it takes to change the course of monetary policy (see Phelps and Taylor 1977; Taylor 1980; Calvo 1983). These papers showed that systematic monetary policy was possible in spite of agents having rational expectations. This was followed by a direct attack on rational expectations, when Phelps and Roman Frydman challenged the idea by demonstrating the implications of each agent having a distinct model of the world in mind (Frydman and Phelps 1983). During the fourth phase, Phelps responded to another challenge, this one that to his natural rate theory presented by the persistent elevation of unemployment in much of the Organisation for Economic Co-operation and Development (OECD) countries in the 1970s, 1980s, and 1990s. He proposed a set of generating general-equilibrium models of equilibrium unemployment (Phelps 1994) that can explain the long swings in the unemployment rate for a given country as well as differences in average unemployment across countries.

Capital Accumulation and Endogenous Growth

One of Phelps's first influential papers was his 'The Golden Rule of Accumulation: A Fable for

Growthmen', which was published in the *American Economic Review* in 1961. In the paper he follows in the footsteps of Ramsey (1928) in deriving the golden rule of capital accumulation that maximizes the long-run level of consumption per capita. According to his golden rule, the savings rate should be set equal to the share of capital in national income. Shortly after writing this paper Phelps went on to introduce the notion of dynamic inefficiency, which is characterized by a state where lowering the rate of saving would raise the utility of all generations, both current and future (Phelps 1965). In contrast, Phelps and Pollack (1968) showed how inefficiently low levels of saving might arise if each generation discounted its own future utility at a lower discount rate than the utility of future generations. This idea later became know as 'hyperbolic preferences' and has been applied to the study of diverse phenomena.

Phelps also introduced many of the ideas that later became a part of what is now known as endogenous growth theory. He went beyond the neoclassical framework and gave people an explicit role in the generation and adoption of ideas. In his 2006 Nobel Prize Lecture he describes the neoclassical growth model in the following words:

> Neoclassical growth theory was conspicuous in having no people in it. It explained the accumulation and investment of physical capital yet the driving force in that story – increases in knowledge, called 'technology' – rains down exogenously, like manna from heaven – and the selection among new technologies is instantaneous, costless and error-free. Nowhere were people required except in the production functions. It would have been better to suppose that machines do all the producing and that people are deployed over the vast range of activities involving management, judgment, insight, intuition and creativity. (Phelps 2006)

Phelps, knowing that technological progress requires people doing research, explicitly modelled technological progress as a function of the number of workers doing research. For constant exponential growth his model calls for an exponential growth of labour inputs into research, which makes the long-run rate of growth of technology ultimately determined by the growth rate of the population. Two implications followed. First, there is a golden rule level of research effort that maximizes the level of consumption per capita, similar to the golden rule of investment. In other words, a society can have excessive research in that consumption per capita is lower than it would be if more people were producing and fewer engaged in research – the gains in technology cannot compensate for lost consumption. The other implication is that a larger population provides a larger number of people doing research and hence makes it possible to climb to a higher technology path. The following quotation is revealing:

> One can hardly imagine, I think, how poor we would be today were it not for the rapid population growth of the past to which we owe the enormous number of technological advances enjoyed today. . . Another instance of external economies is parallel. Our artistic heritage is much like our technology; it is a part of our 'public capital'. If I could re-do the history of the world, halving population size each year from the beginning of time on some random basis, I would not do it for fear of losing Mozart in the process. No improvement of our dirty air and our traffic congestion could compensate me for that! (Phelps 1968b, pp. 511–2)

The adoption of new technology also requires people. Nelson and Phelps (1966) study the implications of managers needing to have an idea about the expected value (net of costs) of a technological innovation and the probability of a successful adoption. They propose the idea that education helps managers in this regard; education enhances the ability to learn, understand and adopt what others have discovered. Accordingly, economic growth in the long run depends on the level of education, not its change, as confirmed by recent empirical work. The Nelson and Phelps paper also introduces the concept of a technology gap between each country and a technology leader, an idea that has become important in recent work on endogenous growth. The steady-state technology gap is shown by Phelps to be a decreasing function of the level of education and a positive function of the rate of change of leading technology.

Inflation and Unemployment

The rejection of a stable inflation–unemployment trade-off is perhaps Phelps's greatest achievement. He did this essentially by bringing expectations into macroeconomic models of inflation and unemployment. By turning expectations into a state variable, reflecting past unemployment/inflation choices, Phelps showed how monetary policy had an intertemporal dimension. By increasing the supply of money and lowering unemployment today inflation is increased, which eventually raises expectations about future inflation and makes the inflation–unemployment trade-off worse – there is no long-run trade-off between unemployment and inflation, contrary to what the economics profession had believed.

In Phelps's 1968 paper in the JPE he sets himself the task of explaining why an increase in the supply of money has a positive effect on output in the short run, instead of just raising prices and wages (Phelps 1968a). The paper provides microeconomic foundations for wage setting, introduces the notion of efficiency wages and equilibrium unemployment, and provides a model of the labour market with job search. Each of these contributions opened up paths for others to research.

In that 1968 paper Phelps models the labour market using a search framework where heterogeneous firms and workers are searching for a suitable match and they meet randomly at a rate determined by the number of unemployed workers searching and the number of vacancies that need to be filled. The frequency of matches is described by a matching function, which makes the paper a forerunner of the matching theory of Diamond, Mortensen and Pissarides. However, as Phelps pointed out later, the existence of unemployment in equilibrium was essentially not dependent on the heterogeneity of workers and labour market search; all that was needed was rising marginal training costs and job heterogeneity that made workers quit their jobs occasionally (see Phelps 1995). In the model, firms and employees have to make their decisions before learning about the decisions made by others. An expectational disequilibrium is created when a positive monetary shock drives unemployment below its equilibrium level and firms experiencing higher quit rates respond by raising their money wages, thinking that this will raise their relative wages. Here Phelps spearheaded the work on efficiency wages, an idea later developed by Steven Salop, Guillermo Calvo, Carl Shapiro and Joseph Stiglitz. But, to continue the present story, observed wage inflation rises when every firm raises its wages and this is soon reflected in expectations of higher wage inflation which makes each firm raise wages even more, hence further increasing actual wage inflation. The only non-inflationary point is at the equilibrium rate of unemployment where expected wage inflation equal actual wage inflation. The paper has the seeds of a model of an endogenous natural rate because the rate of equilibrium unemployment is shown to be a function of the rate of growth of the labour force – an increase in the rate of growth of the labour force raises the level of the equilibrium unemployment rate due to rising marginal costs of hiring.

In a separate paper, Phelps proposed a parable of an economy in which output is produced on separate islands, each having its own labour market. When wages and prices are set on one island, this is done without the knowledge of what is happening on the other islands. When demand goes up, due to a loose monetary policy, individual producers do not realize that this is happening; instead they think this is at least partly caused by the changed preferences of consumers and hence do not raise wages and prices fully to neutralize any output effects. Only gradually do their expectations about prices and wages adjust, making them raise wages further, thus eliminating the output effects (Phelps 1970b).

Phelps treats expectations of wages and prices as a state variable affecting output and employment. What matters for output and unemployment is the deviation of actual wages and prices from their expected values. The implication is that a monetary stimulus has only a short-run effect on employment and output; in the long run both are determined by the structure of the economy (that is, non-monetary factors). This is the *natural-rate hypothesis.* The policy implication that follows is that central banks must be concerned about the effects of their actions on inflationary expectations. If they reduce interest rates today, they may stimulate output and employment but at the cost of higher expected inflation – an upward shift of the short-run Phillips curve – which requires higher interest rates in the future, hence lower employment and output. Monetary policy becomes an intertemporal planning problem (see Phelps 1967, 1972b). This intertemporal dimension of monetary policy is taken quite seriously by independent central banks that target inflation. The intertemporal dimension of policymaking was emphasized by the Nobel committee when explaining its decision to choose Phelps for the prize.

Phelps's treatment of the labour market as plagued by various imperfections and market failures was mirrored in his description of goods markets. Phelps proposed an early model of imperfectly competitive goods markets in a joint paper with Sidney Winter in the Phelps volume of 1970. The basic idea is that consumers have imperfect information about prices and therefore become customers where they believe prices to be lower. However, information about prices gradually spreads between consumers and when a consumer learns about lower prices elsewhere he leaves his present supplier. In this set-up firms treat their market share as an asset, comparable to their stock of capital and trained employees. The markup decision becomes an intertemporal investment decision; a price increase, while raising current profits, gradually causes customers to drift elsewhere, hence reducing future profits. The implication is that the markups decision is affected by macroeconomic variables such as the rate of interest and the expected rate of growth of sales to each customer. A fall in the rate of interest, as well as a rise in expected sales per customer, would make firms cut markups in order to invest in an expanded market share. Similarly, when firms expect an imminent recession they have an incentive to raise prices so that price inflation precedes recessions. The customer market model later played an important role in the general equilibrium models of the natural rate developed in the 1994 book, *Structural Slumps.*

New Keynesian Economics

In the early 1970s Robert Lucas combined Phelps's island parable and the assumption of rational expectations to generate what became known as new classical economics (Lucas 1972). In these models only unexpected demand shocks affect output and employment and, more controversially, the deviations of these variables from their equilibrium values only persist as long as expectations remain incorrect; hence anticipated stabilization policy is ineffective. Phelps responded – often in collaboration with his colleagues at Columbia, John Taylor and Guillermo Calvo – by showing that a firm's expectational errors could have real effects even in models having rational expectations, where there is no lack of understanding or perception of what other firms were up to, because of staggered wage and price contracts. The objective was to establish microfoundations for the Keynesian prediction that a permanent demand shock causes a persistent slump and that monetary stabilization policy can be effective. The proposed models are based on the simple observation that wages and prices are never adjusted continuously, but by convention they are set periodically and the timing of wage and price changes is staggered across firms. However, money is neutral in the long run in the staggering models and output tends towards an equilibrium level, unemployment towards its equilibrium level. Phelps was the first to express the view that a model combining rational expectations and wages and prices being set at regular intervals could give Keynesian results (Phelps 1974). This work later became known as New Keynesian economics (see Phelps and Taylor 1977; Fischer 1977; Taylor 1980; Calvo 1983).

Phelps continued his attacks on new classical economics in the 1980s when, in collaboration with Roman Frydman, he challenged the very notion of rational expectations by expressing scepticism about their relevance when agents' actions depended not only on their beliefs about aggregate variables but also about other agents' beliefs. Individual agents, when acting on their understanding of an economic model, may not converge to a rational-expectations equilibrium because they need to continuously re-estimate while other agents are doing exactly the same. Frydman and Phelps (1983) claim that individual rationality does not guarantee the coordination of beliefs that is assumed in a rational expectations equilibrium, and emphasize the need for a model of learning as an integral part of a model of macroeconomic dynamics.

The Changing Natural Rate

In the late 1980s and early 1990s, Phelps's attention turned to explaining the persistent rise of unemployment in most OECD countries in the previous two decades. While unemployment had been lower in Europe than in the United States in the 1950s and 1960s, European unemployment started its ascent in the 1970s and moved to a higher plateau, where the big continental economies still find themselves. This experience turned out to be a challenge to Phelps's important work on equilibrium unemployment. How come, asked the critics, that unemployment does not revert back to its pre-shock levels? What happened to the natural rate? In spite of Phelps's 1968 *JPE* paper having the seeds of a model of an endogenous natural rate, this model was inadequate when it came to explaining the persistent rise of unemployment from the early 1970s on.

It is a testimony of Phelps's pervasive influence on the theory of unemployment and inflation that initial attempts by others to explain the persistently high unemployment in the 1980s were to a large extent based on ideas taken from his 1972 book, *Inflation Policy and Unemployment Theory.* Here he introduced the concept of 'hysteresis' to economics: there is hysteresis when an equilibrium point depends on the path taken by prices and quantities towards the equilibrium. In the labour market context, the level of equilibrium unemployment may depend on the path taken toward it, that is, a temporary recession may have a permanent effect. In the same book, Phelps went on to consider some possible hysteresis channels. He suggested that unemployment might adversely affect the human capital and work habits of those affected and also that hysteresis could arise due to the dynamics of union membership: a recession reduces the number of union members and this makes those remaining push for higher wages, thus preventing employment from recovering. The hysteresis effects of long-term unemployment working through human capital depreciation were emphasized and developed further by, amongst others, Layard et al. (1991), while Lindbeck and Snower (1988) extended and developed the idea of hysteresis arising from insider–outsider dynamics.

Phelps disagreed with those who believed that hysteresis could explain the failure of unemployment to fall in the 1980s following the steep recessions at the beginning of the decade. He responded with a series of papers that gradually built a general equilibrium model of the determination of equilibrium unemployment – in steady state the natural rate of unemployment. This work culminated in the publication of *Structural Slumps* in 1994. This book has three prototype models that are non-monetary and emphasize the role of various market imperfections in goods and labour markets. The key imperfection in the labour market is asymmetric information, which leads firms to use wages to reduce quitting and shirking. There arises an upward-sloping wage curve in the wage-employment plane that reflects efficiency-wage considerations. Three models described labour demand. The first uses the customer market set-up of Phelps and Winter (1970a) where firms set markups so as to maximize the present discounted value of future profits and customers have imperfect information about prices charged by different suppliers. In this model, a rise in the real rate of interest – or a fall in the expected rate of growth of sales per customer – makes firms disinvest in market share by raising the markup of price over marginal cost, which effectively lowers the real demand wage and raises the natural rate of unemployment. In the second model, firms are concerned about quitting because of the cost of training replacements. Managers use wages to deter quits and the hiring decision also becomes an intertemporal investment decision. Higher interest rates and lower expected productivity growth both make firms reduce hiring as well as lowering wages, which raises the quit rate. The third and final set of models has two sectors: a labour-intensive capital goods sector and a capital-intensive consumer goods sector. A rise in real interest rates makes the relative price of the labour intensive capital good fall, which then lowers the price of labour, raising the natural rate of unemployment. Fitoussi and Phelps (1988) – a precursor to *Structural Slumps* – provide a monetary exposition of some of these effects.

In *Structural Slumps*, as well as in the papers that followed, the elevation of unemployment in the OECD countries in recent decades – particularly on the Continent of Europe – is explained by the simultaneous fall in the rate of productivity growth and the rise of world real interest rates in the early 1980s (see Hoon and Phelps 1997; Phelps and Zoega 1998; Fitoussi et al. 2000). Europe did well in the first two decades following the Second World War because of a combination of low world interest rates and higher productivity growth. Productivity grew at a brisk pace because Europe could imitate the United States – adopt technology that had been developed there in the pre-war years – and meanwhile enjoy high productivity growth in spite of the economic models of the large Continental economies, which otherwise stifled entrepreneurship, initiative and innovation. The closing of the gap and a simultaneous rise in world real interest rates caused a structural slump that monetary policy could not remedy. Analogously, the non-inflationary boom in the United States at the end of the century can be explained by the effect of an anticipated productivity increase in the labour demand wage.

With the passing of time the view that long swings of unemployment require a theory of a changing natural rate of unemployment has gained acceptance. The current debate is focused on the importance of labour market institutions per se and macroeconomic shocks in determining the natural rate of unemployment. Recent work by Phelps has described the adverse effects of Europe's economic model on entrepreneurship, innovation and growth, stemming from its culture as well as the institutions of financial and labour markets, which foster rent seeking and protect vested interests instead of promoting initiative, risk taking and innovation.

Phelps has been interested not only in the macroeconomic causes of unemployment; his interests also extend to the fate of the disadvantaged in modern societies. In the 1990s he wrote a book titled *Rewarding Work* (Phelps 1997) on the problems facing low-skilled workers. This book emphasizes the importance of having a stable job for self-realization, mental stimulation, lending a rhythm to daily life and participation in society as well as income to support one's family and to share in the consumption and leisure activities of others. He describes the worsening of job prospects for the lowest-skilled American workers and proposes a scheme of general subsidies for the lowest paid. The book demonstrates a genuine commitment to help improve society, as do his frequent articles in the *Financial Times* and the *Wall Street Journal*.

Other Contributions

The literature on statistical discrimination originates with Phelps (1972a) and Arrow (1972, 1973). Again we start with asymmetric information, in this case about an individual worker's productivity. Given a statistical correlation between a worker's group attributes and average productivity in the group, an employer may wish to discriminate on the basis of which group the worker belongs to. Unequal treatment of identically productive workers may give a result that does not depend in any way on the employer's preferences or prejudice.

In the field of public finance, Phelps (1973a) found that inflation, being a source of tax revenue, should be chosen optimally along with other forms of taxation. A positive rate of inflation is required to minimize the distortions from different forms of taxation. Finally, there is the 'Phelps–Sadka result', namely, that the marginal tax rate should approach zero at the top of the income distribution because policymakers can observe only wage incomes, not wage rates per hour (Phelps 1973b).

Last but not least, one should mention his *Seven Schools of Macroeconomic Thought* that offers a very personal description of the genesis and distinctive characteristics of the macroeconomics of Keynes, monetarism, the New Classical School, the New Keynesian School, supply-side economics, real business cycle theory, and what he called the Structuralist School, which includes the work done on endogenizing the natural rate of unemployment using nonmonetary models in the 1980s and 1990s.

Concluding Thoughts

This overview of Phelps's work is a testimony to his impact on the history of macroeconomic thought. From the microfoundations of macroeconomics, the attack on the Phillips curve trade-off and equilibrium unemployment, to efficiency wages, optimal monetary policy, staggered contracts and a theory of moving equilibrium rate, Phelps has helped shape our view of the macroeconomy. Moreover, as a person he has clearly inspired and motivated a host of other well-known economists in their work. A surprising number of important contributions trace their origins to his influence. One can say that this particular contribution of his is more subtle yet no less real. For almost half a century Edmund Phelps has contributed to economics, driven by the excitement of discovery and the joys of creativity.

See Also

- ▶ Efficiency Wages
- ▶ Endogenous Growth Theory
- ▶ Golden Rule
- ▶ Microfoundations
- ▶ Natural Rate of Unemployment
- ▶ New Keynesian Macroeconomics
- ▶ Structural Unemployment

Selected Works

1961. The golden rule of accumulation: a fable for growthmen. *American Economic Review* 51: 638–643.

1965. Second essay on the golden rule of accumulation. *American Economic Review* 55: 793–814.

1966. Models of technical progress and the golden rule of research. *Review of Economic Studies* 33: 133–145.

1966. (With R.R. Nelson.) Investment in humans, technological diffusion and economic growth. *American Economic Review* 56: 69–75.

1967. Phillips curves, expectations of inflation and optimal unemployment over time. *Economica* 34: 254–281.

1968. (With R.A. Pollack.) On second-best national saving and game-equilibrium growth. *Review of Economic Studies* 35: 185–199.

1968a. Money wage dynamics and labor market equilibrium. *Journal of Political Economy* 76: 687–711.

1968b. Population increase. *Canadian Journal of Economics* 1: 511–2.

1970a. (With S.G. Winter, Jr.) Optimal price policy under atomistic competition. In Phelps et al. (1970c).

1970b. The new microeconomics in inflation and employment theory. *American Economic Review* 59: 147–60. Revised version in Phelps et al. (1970c).

1970c. (With A.A. Alchian, C.C. Holt, D.T. Mortensen, G.C. Archibald, R.E. Lucas, L.A. Rapping, S.G. Winter, J.P. Gould, D.F. Gordon, A. Hynes, D.A. Nichols, P.J. Taubman, and M. Wilkinson.) *Microeconomic foundations of employment and inflation theory*. New York: W.W. Norton.

1972a. The statistical theory of racism and sexism. *American Economic Review* 62: 659–661.

1972b. *Inflation policy and unemployment theory*. New York: W.W. Norton.

1973a. Inflation in the theory of public finance. *Swedish Journal of Economics* 75: 67–82.

1973b. Taxation of wage income for economic justice. *Quarterly Journal of Economics* 87: 331–354.

1974. Remarks on monetary policy-making under rational expectations. Unpublished paper, conference on rational expectations and monetary policy. Minneapolis: Federal Reserve Bank of Minneapolis.

1977. (With J.B. Taylor.) Stabilizing powers of monetary policy under rational expectations. *Journal of Political Economy* 85: 163–90.

1983. (With R. Frydman.) *Individual forecasting and aggregate outcomes: 'Rational expectations' examined*. Cambridge: Cambridge University Press.

1988. (With J.P. Fitoussi.) *The slump in Europe: Open economy theory reconstructed*. Oxford: Basil Blackwell.

1990. *Seven schools of macroeconomic thought.* Oxford: Clarendon Press.

1994. *Structural slumps: The modern-equilibrium theory of unemployment, interest and assets*. Cambridge, MA: Harvard University Press.

1995. The origins and further development of the natural rate of unemployment. In *The natural rate of unemployment: Reflections on 25 years of the hypothesis,* ed. R. Cross. Cambridge: Cambridge University Press.

1997. *Rewarding work: How to restore participation and self-support to free enterprise*. Cambridge, MA: Harvard University Press.

1997. (With H.T. Hoon.) Growth, wealth and the natural rate: Is Europe's jobs crisis a growth crisis? *European Economic Review* 41: 548–557.

1998. (With G. Zoega.) Natural-rate theory and OECD unemployment. *Economic Journal* 108: 782–801.

2000. (With J.P. Fitoussi, D. Jestaz, and G. Zoega.) Roots of the recent recoveries: labor reforms or private sector forces. *Brookings papers on economic activity* (1): 237–91.

2006. My kind of macroeconomics: modern economies and their policy choices.

Prize Lecture, Nobel Prize in Economics. 2006, 8 December. Online. Available at: http://nobelprize.org/nobel_prizes/economics/laureates/2006/phelps-lecture.html. Accessed 5 June 2007.

Bibliography

Aghion, P., R. Frydman, J.E. Stiglitz, and M. Woodford. 2003. *Knowledge, information and expectations in modern macroeconomics*. Princeton: Princeton University Press.

Arrow, K.J. 1972. Models of job discrimination. In *Racial discrimination in economic life*, ed. A.H. Pascal. Lexington: D.C. Heath.

Arrow, K.J. 1973. The theory of discrimination. In *Discrimination in labor markets*, ed. O. Ashenfelter and A. Rees. Princeton: Princeton University Press.

Calvo, G.A. 1983. Staggered prices in a utility-maximizing framework. *Journal of Monetary Economics* 12: 383–398.

Fischer, S. 1977. Long-term contracts, rational expectations, and the optimal money supply rule. *Journal of Political Economy* 85: 191–205.

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

Layard, R., S. Nickell, and R. Jackman. 1991. *Unemployment: Macroeconomic performance and the labour market*. Oxford: Oxford University Press.

Lindbeck, A., and D. Snower. 1988. *The insider–outsider theory of employment and unemployment*. Cambridge, MA: MIT Press.

Lucas, R.E. 1972. Expectations and the neutrality of money. *Journal of Economic Theory* 4: 103–124.

Ramsey, F. 1928. A mathematical theory of saving. *Economic Journal* 38: 543–559.

Taylor, J.B. 1980. Aggregate dynamics and staggered contracts. *Journal of Political Economy* 88: 1–23.

Philippovich von Philippsberg, Eugen (1858–1917)

H. C. Recktenwald

Abstract

A Viennese by birth, Philippovich began his academic career as a Lecturer (*Privatdozent*) in Vienna (1884), after studies in economics in Graz, Vienna and Berlin. He obtained a chair in Freiburg (Breisgau) and returned to Vienna (1893), where he worked until his death. Deeply interested in the economic problems of his time and passionately engaged in social reforms, he took an active part in Austrian and German political life. As a member of the upper house of parliament and the leading spirit of the 'Austrian Fabians', Philippovich had a significant influence on social legislation in Austria. Like the German academic socialists (*Kathedersozialisten*) Schmoller and Wagner, members of the Verein für Socialpolitik (of a New Deal type) with whom he cooperated closely, he was a regulationist of the market process, who attributed to the state a moral and economic competence which seems to be wishful thinking rather than based on reason and experience. He strongly believed that a 'middle course' between socialism and a competitive economy would help to ease economic and social tensions which were largely caused by overpopulation along with the overcrowded labour market during Germany's transition from an agrarian into an industrial nation. On the other hand he realized better than most of the German reformers that a sound analytical foundation was necessary for any policy of reasonable reforms.

A Viennese by birth, Philippovich began his academic career as a Lecturer (*Privatdozent*) in Vienna (1884), after studies in economics in Graz, Vienna and Berlin. He obtained a chair in Freiburg (Breisgau) and returned to Vienna (1893), where he worked until his death. Deeply interested in the economic problems of his time and passionately engaged in social reforms, he took an active part in Austrian and German political life. As a member of the upper house of parliament and the leading spirit of the 'Austrian Fabians', Philippovich had a significant influence on social legislation in Austria. Like the German academic socialists (*Kathedersozialisten*) Schmoller and Wagner, members of the Verein für Socialpolitik (of a New Deal type) with whom he cooperated closely, he was a regulationist of the market process, who attributed to the state a moral and economic competence which seems to be wishful thinking rather than based on reason and experience. He strongly believed that a 'middle course' between socialism and a competitive economy would help to ease economic and social tensions which were largely caused by overpopulation along with the overcrowded labour market during Germany's transition from an agrarian into an industrial nation. On the other hand he realized better than most of the German reformers that a sound analytical foundation was necessary for any policy of reasonable reforms.

Evidently his interest in theory increased (*Die Entwicklung,* 1910) and was influenced by the Austrian school of Menger, Wieser and Böhm-Bawerk. Indeed, in his *Grundriss* (1893–1907), a leading German textbook on economics for a whole generation, he successfully attempted to bridge the gap between the two opposing schools: it was mainly through this channel that Austrian theories and marginal utility analysis reached German students and not via Gossen's or even von Thünen's pioneering works.

Like many of his contemporaries, Philippovich was a careful thinker and a great teacher of intellectual stature who sometimes liked to mistake the chair for the pulpit in order to preach instead of explain and reason.

See Also

- ► Austrian Economics

Selected Works

1885. *Die Bank von England in Dienste der Finanzverwaltung des Staates*. Vienna: Toeplitz and Deuticke. 2nd revised edn, 1911, trans. by C. *Meredith as A history of the Bank of England, and its financial services to the state*, Washington: National Monetary Commission, 1911.

1886. *Über Aufgabe und Methode der politischen Ökonomie*. Freiburg: Mohr.

1893–1907, 1914–16. *Grundriss der politischen Ökonomie,* 2 vols. Freiburg/Tübingen: Mohr.

1910. *Die Entwicklung der wirtschaftspolitischen Ideen im neunzehnten Jahrhundert*. Tübingen: Mohr.

Bibliography

Recktenwald, H.C., eds. 1973. *Political economy: A historical perspective*. London: Collier-Macmillan.

Stigler, G.J. 1982. *The economist as Preacher and other essays*. Chicago: University of Chicago Press.

Phillips Curve

Edmund S. Phelps

Abstract

In 1957 A.W. Phillips argued that, other things equal, the rate at which the nominal wage level was changing was a decreasing function of the level of the unemployment rate. Further, the rate of unemployment required to keep the rate of wage inflation down to the normal level was certainly positive in the United Kingdom, the domain of Phillips's data, had remained stable for nearly a century. Milton Friedman and Edmund Phelps criticized the concept of a stable Phillips curve for having treated wage-setters' behaviour, which presumably involved their expectations of the general wage movement, as a mechanical toy.

Keywords

Aggregate demand; Cost inflation; Equilibrium price level; Excess demand and supply; Fellner, W. J.; Fisher, I.; Friedman, M.; Hyperinflation; Inflation; Keynes, J. M.; Labour supply; Lipsey, R. G.; Lucas, R.; Monetary theory; Muth, J.; Natural rate of unemployment; Nominal wages; Phelps, E. S.; Phillips curve; Phillips, A. W.; Political business cycles; Rational expectations; Samuelson, P. A.; Sargent, T.; Unemployment; Wage inflation

JEL Classifications

E1

By the 1950s there was achieved a working synthesis, despite some unsolved problems, of the contributions of Keynes to monetary theory with the older truths of his several predecessors Marshall, Pigou, Wicksell and Fisher. Given the supply of money, there is a nominal price level that is in some suitable sense the equilibrium price level; more generally, there is an equilibrium path of the price level. The equilibrium price level in the current period, given next period's price level, is just high enough to reduce the real value of this period's cash balances down to the quantity demanded – figured at the corresponding nominal rate of interest (which is a decreasing function of this period's price level) and output level (which was taken to be independent of the price level if nominal wages were also taken as finding their equilibrium level). If people expect that the general level of prices and nominal wages is higher, and we assume that the actual price level at first equals this expected level, the result will be disappointment – an unexpected weakening of sales. Presumably, the price and wage levels will then tend to adjust, and perhaps employment will detour from its equilibrium level in the process.

The disequilibrium dynamics of the adjustment process, however, remained *terra incognita*. Suppose that is a sudden and unexpected disturbance that displaces upwards or downwards the path of the equilibrium price level. Keynes had declared in his 1936 book that the money wages set by producers would not generally take the downward jumps occasionally necessary for continued maintenance of equilibrium, hence the need for a more general theory of interest and employment in which the nominal wage level was not on the equilibrium track. (He further opined that lessened wage inflexibility would be destabilizing.) By the 1950s it was agreed that wages would *gradually* move from the former equilibrium path, if we assume they were originally in equilibrium, toward the new and lower equilibrium path, whether or not there would be later overshooting, and further that, if there is such gradualness, the result will be a bulge of unemployment during the process of wage adjustment. Similarly, an upward displacement of the equilibrium path would likewise engender only a gradual adjustment of money wages, accompanied in this case by a dip of the unemployment rate below its equilibrium, or normal, level. Increasingly, economists spoke of buying a spell of abnormally low unemployment by generating a round of inflation. (Yet, some economists of Austro-Hungarian or German schooling, notably William Fellner, argued that successive doses of (equal) inflation would lose their effectiveness, so that the same effect on unemployment would require ever increasing doses, as anticipations of higher demand came to be built into wage contract increases.) The term cost inflation arose to refer to the sort of inflation the avoidance of which needed the discipline, and social waste, of unemployment above what could be achieved through high demand.

It was against this background that A.W. Phillips's extraordinary article, scholarly yet accessible, appeared in the academic journal *Economica* in 1957. Phillips changed the terms of discourse of the subject from the qualitative and discontinuous to ordinary quantitative terms. Other things equal, such as the rate of change of unemployment, the rate at which the nominal wage level is changing – the (algebraic) rate of wage inflation – is a decreasing function of the level of the unemployment rate. Further, the rate of unemployment required to hold down the rate of wage inflation to the level of normal experience – the average, and accustomed, rate – is certainly positive, perhaps 2 to 3 per cent in the United Kingdom, the domain of Phillips's data, and has not shifted notably over nearly a century of observation. Almost overnight the Phillips curve (so named in a discussion by Samuelson and Solow) invaded the language of macroeconomics.

Phillips uncovered another fact about past wage inflation. Among years with the same (annual) level of the unemployment rate there tended to be a higher rate of wage inflation when the annual unemployment rate was falling, as in a cyclical recovery or developing boom, than when the annual unemployment rate was rising. Phillips drew a counterclockwise loop around the downward-sloping Phillips curve, to indicate the typical motion of the wage inflation rate in relation to the unemployment rate over the typical historical cycle. (See the lower Phillips curve and the loop around it in Fig. 1.) It remained for R.G. Lipsey, also of the London School of Economics at that time, to express this historical phenomenon in quantitative terms too. Lipsey in 1960 published estimates obtained by regression analysis of the coefficients of a linear rate-of-wage-change equation in which the explanatory righthand-side variables were the level of the unemployment rate and its rate of change. The negative sign of Lipsey's estimate of the latter coefficient reflected the above loop. The statistical estimation of such Phillips–Lipsey equations rapidly developed from a cottage activity using electric calculators to a booming computerized industry.

In a way, the new and developing fact book seemed to contain information that was entirely reasonable and surely in keeping with existing theoretical (or pretheoretical) notions. (Indeed, a remarkably early anticipation of the Phillips curve was later unearthed in an obscurely placed paper in 1926 by Irving Fisher.) It seemed to say, essentially, that if there was an aggregate excess supply then nominal wages would be found falling and employment would be depressed – as long as wages remained too high to eliminate the excess supply – and both effects of the excess supply would be larger the greater was the size of the excess supply. More exactly, the sudden appearance of an excess supply that is maintained at a given level for a while would first generate a positive rate of change of unemployment alongside falling wages and only later, in a sort of disequilibrium steady state, a higher level of the unemployment rate without a positive rate of change. This part of the Phillips curve story seemed unsurprising and unpuzzling.

Yet some theoretical problems that had long lain submerged and unnoticed when the subject of disequilibrium adjustment was still muddy and relatively quiet came to surface once the Phillips–Lipsey formulation had stirred things up. Among these was the problem of explaining why nominal wages did not jump down to their new equilibrium level (with prices jumping after them) and, beyond that, the problem of determining the pace with which wages fell. The same theoretical void had been created more than a decade before Phillips's article when Samuelson in his *Foundations*, addressing Walrasian stability, simply postulated that the rate at which the price of a commodity falls is an increasing function of the excess supply of it. This was a macroeconomic hypothesis, perhaps a kind of theory by the behavioural standards of the day, but not a microeconomic theory running in terms of the motives and perceptions of the individual actors operating in the economy.

If the first problem was explaining that the Phillips curve was sloping, the second problem was explaining its remarkably rightward position: Money wage rates tended to be rising over a range of positive unemployment rates, including rates exceeding the lower bound obtainable by high-pressure aggregate demand levels. If nominal wages tend to be rising as long as the unemployment rate stays above nondepression levels, then the Samuelsonian hypothesis explains that markets have normally operated in a state of considerable excess demand. But is that likely? Is a state of zero excess demand (and excess supply) really marked by a zero rate of wage change, or is something missing here? Somehow, it was evident, the factors of productivity growth and inflation needed to be brought into the analysis, but not just as incantations to make the problem go away.

For many economists there was the further problem of reconciling the empirical regularity depicted by the Phillips curve, which seemed to possess an extraordinary stability, with the older Continental, or Austro-Hungarian, doctrine, propounded by Fellner and others, that below-normal unemployment constantly fuelled by a permissive monetary–fiscal policy will soon cause wages (and hence prices) to rise in ever-accelerating fashion until the hyperinflation finally brings collapse or structural change. This was the further problem of understanding in microeconomic terms the shiftability of the Phillips curve.

The solution of the first problem, that of explaining the gradualness of the wage adjustment and the attendant slump of employment, led theorists in the 1960s in the same direction in which Keynes had been led in his search for an explanation of slumps. A key element of the solution was the fact that there is no coordination, to use Keynes's term, among the managers deciding upon wages and employment (inter alia) at the various production sites. If there is a weakening of aggregate demand – here, a curve in the output–price level plane – in a previously normal and equilibrium situation, the resulting fall in the demand curve facing the individual manager, or producer, even if seen by him as permanent, would not induce the workers employed there (or unemployed there) to accept the job-preserving money wage cut unless they were expecting workers elsewhere at the same moment to be facing and accepting the very same percentage wage cut; and they would have not reason to have that expectation unless there was news bearing on the scale of the decline in demand and such news was observed to have produced job-preserving wage cuts. Pending such news, then, there would be only an insufficient wage cut, so the supply price of output would fall by less than the demand price, and hence output and employment would decline. These impact effects would show a negative correlation between wage change and unemployment level (though here the true correlation is with the change of employment).

In the 1960s, however, a number of theorists pointed out the theoretical existence of a deeper Phillips curve relation. The higher unemployment level comes about because 'expected wages' in the economy as a whole exceed 'actual wages', and as information comes in that actual wages elsewhere are lower than expected the ensuing downward revision of expectations will induce workers to accept still lower actual wages. This latter wage fall grows out of the disequilibrium situation, like the higher unemployment. If one were to go so far as to posit static expectations, so that each observed wage decline is thought to be the last, there would exist a disequilibrium steady-state relationship between the size of the (swelling of the) unemployment rate and the magnitude of the rate of wage change. A 1969 Pennsylvania conference developed these points in a variety of models, and the conference volume published a year later served to popularize these expectational microeconomic foundations of unemployment and wage-price behaviour (Phelps et al. 1970).

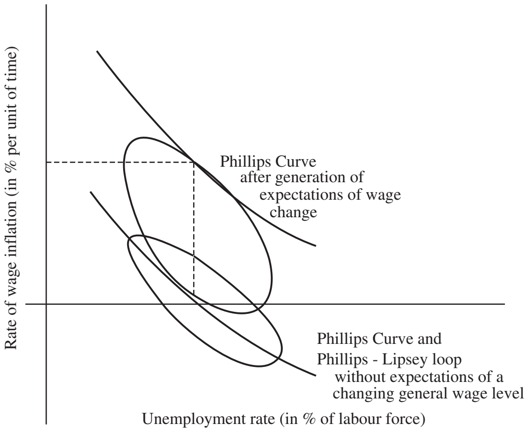
In the 1970s theorists moved toward rational expectations in the sense of Muth. In this case, the news of the initial fall of wages (together with any news on the unemployment front) is enough for workers to expect that the general wage level will now fall to exactly the job-preserving level, so that the unemployment rate will return to the equilibrium level; otherwise workers are implied to be repeatedly misforecasting the wage level, contrary to rational expectations. Here, too, the high unemployment precedes a wage fall (though large enough to eliminate the high unemployment), so that there is again a negative correlation between unemployment level and wage change. A microtheoretic model along these lines, involving known stationary stochastic processes, was developed by R.E. Lucas (1972, 1973) and an intertemporal model with which to show, as a corollary, the ineffectiveness of preannounced monetary policy in stabilizing output or employment was analysed by T.J. Sargent (1973).

The rational expectations postulate seemed at first to point to the conclusion that, following an unexpected drop of aggregate demand, nominal wages would indeed jump – though too late to prevent a recession – once the news of the economic indicators signalling a slump was out, and that with that jump the unemployment rate would jump back to its steady-state equilibrium (and normal) level. But that would have been jumping to conclusions, and fortunately so for the rational expectations hypothesis since there is convincing econometric evidence that the unemployment rate displays statistical persistence. It was soon remembered, however, that the antecedent literature on the costs of recruitment or training provided the basis for an equilibrium path of recovery from a downturn along which both the unemployment rate and the nominal wage level decline continuously, or gradually in discrete-time terms, rather than with a jump. There was also a development of the point made in the earlier literature that firms' wage commitments are apt to be durable and non-synchronous, so that the respective firms in the economy take turns over the wage-setting cycle, or 'year', in resetting their 'annual' wage scales. In such a nonsynchronous wage-setting context, the average level of nominal wages cannot jump and hence employment will not recover from a recession with a jump. Further, a model of wage staggering, though quite different from the preceding types, likewise produces an explanation of the negative correlation between wage change and the unemployment rate, as shown by Taylor (1980).

The second Phillipsian problem, that of explaining the coexistence of rising nominal wages with above-minimum unemployment, had two answers, independent and additive. One answer lay in divorcing ourselves from thinking of the unemployment rate – or even the excess of the unemployment rate over the minimum rate achievable by stimulating aggregate demand – as a satisfactory measure of downward pressure on nominal wages. If the unemployment rate (or, more accurately, the aforementioned excess rate) were driven to zero, quitting would presumably be rampant and so the representative firm would endeavour to pay a wage premium – a positive differential over the wages paid elsewhere. If this average wage level is expected to be unchanged, the firm will therefore raise its wage to a level in excess of that average, with the consequence that the average wage will actually rise – resulting in an excess of 'actual' over 'expected', thus a disequilibrium. It is only when the unemployment rate (or the excess rate) is positive and high enough that the quit rate will be damped sufficiently to encourage the representative firm to content itself with paying the representative wage, that the average wage will remain flat as expected. (The argument is implicit in Phelps 1968, and the explicit focus of Stiglitz 1974, and Salop 1979.) In this equilibrium there is involuntary unemployment in a natural sense of the term, since wages exceed the market-clearing level, and this unemployment may very well exceed job vacancies (if any), so there may be considerable excess supply. (See also Calvo 1979, for another model.)

The other answer to the problem lay in realizing that wages do not rise only when firms (or at least the representative firms) want to be more competitive than the others. Wages may also rise because the firms believe they must raise their wages just to avoid losing any of their present competitiveness. The same point can be made in terms of the excess-demand framework of Samuelson: the error in Samuelson's formulation was in excluding the possibility that wages will be increased in an anticipatory move that serves to prevent the emergence of an excess demand, not just in response to excess demands that are not previously expected and forestalled by intervening wage increases. Hence, nominal wages may be rising not because the labour market is in disequilibrium, marked by mutually inconsistent desires among the firms for superior competitiveness in the labour market, but rather because the prospect of productivity growth or of inflation or of both generates expectations that the general level of wages is going to increase (Phelps 1968).

With the latter insight our third problem, that of explaining the possible shift of the Phillips curve, is also solved. When governments seek to exploit the Phillips curve by trading off price stability in hopes



Phillips Curve, Fig. 1 Wage and unemployment dynamics

of obtaining reduced unemployment in return, they ultimately engender expectations of regularly increasing wages. Such an increase in the expected rate of wage inflation (at each level of the unemployment rate) shifts up the Phillips curve; a new one arises corresponding to the new expected rate of wage inflation. In Fig. 1 see the upper Phillips curve, which has been driven higher by expectations of a rising general wage level. It is now evident that a political business cycle, by alternately lifting and depressing the Phillips curve, would generate the clockwise loop shown in the figure.

If we posit, as a plausible approximation, that the expected wage inflation variable takes its place among the explanatory right-hand variables (alongside the. Phillips–Lipsey terms) with a unitary coefficient, the implication is that the steadystate equilibrium unemployment rate – at which expectations are borne out – is the same number independently of the inflation rate. Then, maintaining a steady unemployment rate below that constant equilibrium rate would entail rising inflation without bound (Phelps 1968; see also Friedman 1968, discussed below). With this coefficient value of one (or any larger value) the model gives algebraic expression to the abiding accelerationist fears of the Austro-Hungarian school.

The notion that the equilibrium unemployment rate was a constant, as above, also emerged from a quite different formulation by Friedman (1968), where the constant was dubbed the natural rate of unemployment. There the rate of wage change is postulated to be a function of the unemployment rate plus the expected rate of *price* inflation. The implicit rationale was that the amount of labour supplied as an increasing function of the expected real value of the nominal wage. A way to synthesize the above wage–wage model (in which expected real-wage changes are captured in the Phillips–Lipsey terms) with the wage–price model is to add to a quasi-Phillips *employment* term a weighted average of the expected rates of wage inflation and price inflation where the latter weight is positive, zero, or negative as the labour supply curve is forward rising, vertical, or backward sloped (Phelps 1979).

See Also

- ► Neoclassical Synthesis

Bibliography

Calvo, G.A. 1979. Quasi-Walrasian theories of unemployment. *American Economic Review* 69: 102–107.

Fellner, W.J. 1959. Demand inflation, cost inflation and collective bargaining. In *The public stake in union power*, ed. P.D. Bradley. Charlottesville: University of Virginia Press.

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

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

Lipsey, R.G. 1960. The relation between unemployment and the rate of change of money wage rates: A further analysis. *Economica* 27: 1–31.

Lucas, R.E. Jr. 1972. Expectations and the neutrality of money. *Journal of Economic Theory* 4 (2): 103–124.

Lucas, R.E. Jr. 1973. Some international evidence on output–inflation tradeoffs. *American Economic Review* 63: 326–334.

Phelps, E.S. 1968. Money-wage dynamics and labor-market equilibrium. *Journal of Political Economy* 76(4): 678–711, Pt II, July–August.

Phelps, E.S. 1979. *Studies in Macroeconomic Theory. Vol. 1: Employment and Inflation*. New York: Academic Press.

Phelps, E.S., et al., eds. 1970. *Microeconomic foundations of employment and inflation theory.* New York: Norton.

Phillips, A.W. 1958. The relation between unemployment and the rate of change in money wage rates in the United Kingdom 1861–1957. *Economica* 25: 283–299.

Salop, S.C. 1979. A model of the natural rate of unemployment. *American Economic Review* 69: 117–125.

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

Samuelson, P.A., and R.M. Solow. 1960. Analytical aspects of anti-inflation policy. *American Economic Review, Papers and Proceedings* 50: 177–194.

Sargent, T.J. 1973. Rational expectations, the real rate of interest, and the natural rate of unemployment. *Brookings Papers on Economic Activity* 1973 (2): 429–472.

Stiglitz, J.E. 1974. Alternative theories of wage determination and unemployment in LDC's: The labor turnover model. *Quarterly Journal of Economics* 88 (2): 194–227.

Taylor, J.B. 1980. Aggregate dynamics and staggered contracts. *Journal of Political Economy* 88: 1–23.

Phillips Curve (New Views)

Jonas D. M. Fisher

Abstract

A *Phillips curve* is an equation which relates the unemployment rate, or some other measure of aggregate economic activity, to a measure of the inflation rate. Since there is a significant correlation between inflation and unemployment over some horizons, understanding this correlation should yield insight into the impulses the economy faces and the mechanisms that propagate their effects. Since the 1990s, research has focused on making progress in three main areas: forecasting, microeconomic foundations and empirical tests of the microfoundations.

Keywords

Business cycles; Cobb–Douglas functions; Expectations-augmented' Phillips curve; Inflation; Inflation forecasting; Labour's share of GDP; Menu costs; Microfoundations; Phillips curve; Price indexation; State-dependent models; Sticky prices; Technology shocks; Unemployment

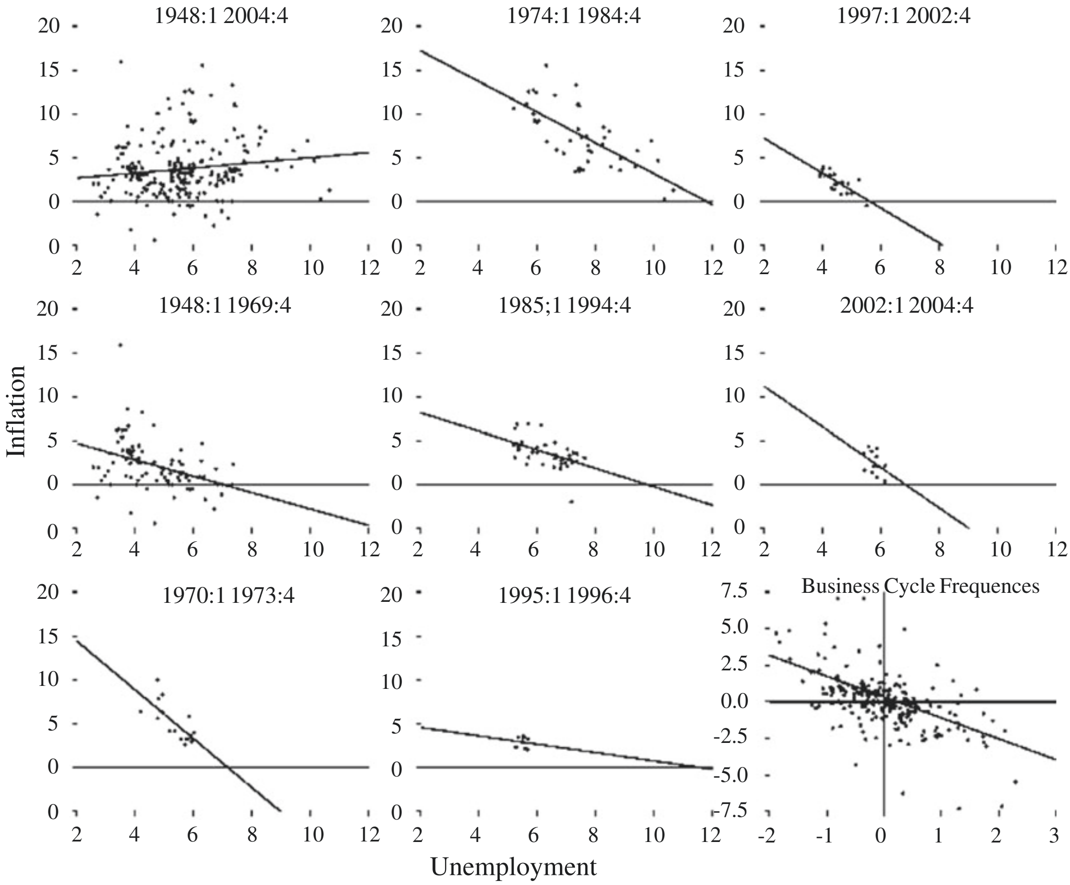
JEL Classifications

D4; D10

A *Phillips curve* is an equation which relates the unemployment rate, or some other measure of aggregate economic activity, to a measure of the inflation rate. This equation continues to prompt a lot of research in macroeconomics, as it has for most of the years since the influential Phillips (1958) and Samuelson and Solow (1960) articles. The early work documents a negative relationship between the unemployment rate and either nominal wage growth or inflation. Equations relating the unemployment rate to inflation were the first to be called Phillips curves. Samuelson and Solow (1960) were bold enough to posit a stable and exploitable structural relationship between unemployment and inflation. The viability of a policy of using inflation to combat unemployment was debunked theoretically in Friedman's (1968) classic presidential address and empirically in the subsequent decade.

The rise of inflation over the 1970s came along with a breakdown in the inflation unemployment relationship and gave birth to the 'expectations-augmented' Phillips curve. This formulation allows the relationship between unemployment to shift due to changes in inflation expectations. Figure 1 shows how such a formulation can be used to fit the data. This shows scatter plots of unemployment and NIPA personal consumption deflator inflation for different sub-periods over the years 1948–2004 along with regression lines. Table 1 reports the regression coefficients, R^2 for the regressions and the means of inflation and unemployment. For the whole sample there is a significant *positive* relationship. However there is always a sequence of consecutive dates where the regression line is negative. The slope coefficient is also highly significant in all cases but one. The movements in the regression line occur as changes in the mean inflation and unemployment rates. Another way to verify that there is a strong association between inflation and unemployment is to focus on business cycle frequencies, in the bottom right hand corner of Fig. 1 and the second row of Table 1. Clearly inflation and unemployment are highly correlated at business cycle frequencies.

Since there is a significant correlation between inflation and unemployment over some horizons, understanding this correlation should yield insight



Phillips Curve (New Views), Fig. 1 The US Phillips curve, 1948:1–2004:4 (*Sources*: authors' calculations; unemployment rate: US Bureau of Labor Statistics; personal consumption expenditure deflator: US Department of Commerce)

Phillips Curve (New Views), Table 1 The US Phillips curve, 1948:1–2004: 4

Sample	Intercept	Slope	R^2	Mean inflation	Mean unemployment
Full sample	2.06***	0.29**	0.02	3.70	5.64
Business cycle frequencies	0.32****	****	0.39	0.00	0.00
1948:1–1969:4	6.55****	****	0.16	2.20	4.67
1970:1–1973:4	20.0****	****	0.52	5.07	5.35
1974:1–1984:4	20.7****	****	0.49	7.54	7.49
1985:1–1994:4	10.2****	****	0.27	3.52	6.41
1995:1–1996:4	5.51	–0.48	0.01	2.89	5.50
1997:1–2001:4	11.1****	****	0.56	2.24	4.47
2002:1–2004:4	15.7****	*	0.17	2.46	5.57

Note: The number of asterisks from one to four denotes significance at the 10, 5, 1, and 0.1 per cent levels of the constant and slope terms of a regression of inflation on unemployment. Business cycle frequencies means the data have been subjected to Christiano and Fitzgerald's (2003) band pass focusing on a 2–8 year horizon. *Sources*: authors' calculations; unemployment rate: US Bureau of Labor Statistics; personal consumption expenditure deflator: US Department of Commerce into the impulses the economy faces and the mechanisms that propagate their effects. Since the 1990s, research has focused on making progress in three main areas: forecasting, microeconomic foundations and empirical tests of the microfoundations. This article reviews the recent research in each of these areas.

Forecasting with the Phillips Curve

Inflation forecasting models rely heavily on the Phillips curve. For many years, even as the traditional Phillips curve relationship evaporated, variables such as the unemployment rate have continued to be very useful predictors of future inflation. Stock and Watson (1999) argued they could do better. They proposed using principal components of large numbers of data series to aid in forecasting macroeconomic variables. The idea was that this approach uses the information in a large number of variables, which is impossible with traditional regression-based forecasting. One of their most interesting findings involves the first principal component of roughly 80 macroeconomic variables, including measures of production and income, employment, unemployment and hours, personal consumption and housing, and sales, orders and inventories. They argued that this 'activity index' variable is more useful than even unemployment for predicting inflation. Such a finding strongly suggests a connection between current activity and future inflation, essentially the Phillips curve relationship.

Atkeson and Ohanian (2001) argued that the success of the Phillips curve in forecasting is just as illusory as a stable Phillips curve. They argued that, for forecasting one year ahead, a simple random walk suffices – the best predictor of one-year-ahead inflation is current inflation. Atkeson and Ohanian's (2001) finding has proven to be remarkably robust (see Brave and Fisher 2004; Fisher et al. 2002). However, the random walk result does depend on the sample period considered by Atkeson and Ohanian (2001), which is 1984–99. Beginning the sample in 1984 is justified by evidence of a major structural change around that time (see Fisher 2006). However, as the sample is extended, the random walk loses some of its lustre. The poor performance of Phillips curve-based forecasting models is mainly confined to the period 1984–93. Since the mid-1990s the traditional variables such as unemployment have been useful forecasters. These findings are easily explained by noting that inflation was generally falling from 1984 to the mid-1990s as the economy adjusted to the Federal Reserve's stronger willingness to fight inflation. It is natural for old models to fail after a major structural change. Moreover, in an environment where output is growing strongly while inflation is falling, it is not surprising the random walk model does well between 1984 and 1993.

Microfoundations of the Phillips Curve

Since Lucas (1972) economists have known how to formulate models in which inflation and activity are correlated but there is not a policy-exploitable Phillips curve. The focus of much of the recent literature has been on the Calvo–Yun Phillips curve, which arises from one particular model. The Phillips curve in this model is named after Calvo (1983) and Yun (1996). Most of the literature uses the hopelessly ambiguous term 'New Keynesian' to describe this model of the Phillips curve.

Phillips curves arise naturally in models where firms set prices and at least some of those prices do not respond to every shock to the economy. Calvo's contribution is a very simple model of sticky prices. He assumed monopolistically competitive firms could re-optimize their price with a fixed probability, θ, each period so that firms re-optimize prices on average every $1/(1 - \theta)$ periods (usually quarters of a year). This formulation can be taken literally, in which case prices are fixed until the next opportunity to re-optimize. Alternatively, firms might follow simple pricing rules at high frequencies and occasionally adjust these rules overtime. Under this interpretation firms can index their prices to inflation.

Yun derived a Phillips curve by introducing the Calvo model of price adjustment into an otherwise standard monetary model with monopolistic competition and constant markups. The result is the Calvo–Yun Phillips curve:

$$\hat{\pi}_t = \beta E_t \hat{\pi}_{t+1} + \frac{(1-\beta\theta)(1-\theta)}{\theta} AD\hat{s}_t. \quad (1)$$

The variable $\hat{\pi}_t$ is the deviation of the log of the gross inflation rate from its steady state value, $\hat{s}_t$ is the log deviation of real marginal cost for the representative firm, and β is the time discount factor of the representative household. The A and D terms are equal to unity in the Yun paper. This equation is derived from the log-linearized necessary conditions of the equilibrium. To linearize around a steady state with positive inflation, firms must index their prices to inflation.

Eichenbaum and Fisher (2007) describe how Kimball's (1995) extension to variable markups implies that $0 < A \leq 1$, where A depends on the shape of the firm's demand curve and equals unity in the constant markup case. Eichenbaum and Fisher also study Woodford's (2003, 2005) model of capital adjustment and describe how this yields.

$0 < D \leq 1$ where D depends on the firm's supply curve. Generally, marginal cost is increasing in output. Since β and θ also lie between zero and unity, the coefficient in front of marginal cost is positive and (1) is an equilibrium relationship where output and inflation are positively related. In most of the literature assumptions are such that $A = D = 1$. This literature generally predicts reasonably large effects of monetary shocks if firms adjust their prices once a year.

The Calvo model is called a *time-dependent* model because the opportunity to change prices depends only on the passage of time. Taylor's (1980) model where firms rotate changing their prices is also a time-dependent model. The main alternative is *state-dependent* models, where changing the price is a choice of the firm which depends on both firm-level variables such as productivity and aggregate variables like the interest rate. The dominant state-dependent model involves menu costs. Studying state-dependent models is more difficult than time-dependent models because the price distribution is endogenous.

Five papers make major progress toward understanding menu cost models. Dotsey et al. (1999) study a model with random menu costs and Taylor-style staggering. A key advantage of their model is that it can be linearized like a simple real business cycle model. Klenow and Krsystov (Klenow and Krystov 2005) calibrate this model to US consumer price index (CPI) micro data for the years 1988–2003. They find that matching the micro data yields a model which behaves very much like the Calvo–Yun model. Golosov and Lucas (2003) study a menu cost model with a constant menu cost but where firms face exogenous technology and/or preference shocks. Under the assumption that the shocks are Gaussian, Golosov and Lucas find that firms choose to adjust their prices a lot when there is a monetary shock, and this makes prices flexible enough that monetary shocks have small affects. Midrigan (2005) uses scanner data to determine the distribution of technology or preference shocks in the Golosov–Lucas model. He estimates this distribution to be non-Gaussian with fat tails. With the estimated distribution monetary shocks have affects similar to models with a Calvo–Yun Phillips curve. Gertler and Leahy (2005) develop an analytically tractable state-dependent model which also behaves like the Calvo–Yun model.

Another key area of research involves building fully specified dynamic general equilibrium models with Phillips curves which fit the data well. This work has focused on the Calvo–Yun Phillips curve instead of more deeply motivated models because of its simplicity. The key contribution is Christiano et al. (2005). Their model also includes portfolio rigidities, adjustment costs in capital, and a Calvo-style version of nominal wage setting. They find their model does a good job matching the evidence on how the economy responds to a monetary shock, with a small amount of price stickiness, but wages must be more rigid. There is a growing amount of research which reaches the same basic conclusion that the wage–activity relationship is more important for understanding macroeconomic dynamics than the traditional Phillips curve (cf. Galí et al. 2007).

Empirical Evaluation of the Microfoundations

Equation (1) is to the empirical macro literature as the Lucas (1978) asset pricing relationship is to empirical finance. In recent years it has come under considerable empirical scrutiny. Galí and Gertler (1999) were the first to use Hansen's (1982) generalized method of moments (GMM) to estimate θ and test (1). They measured marginal cost using labour's share of GDP, which is true if firms use a Cobb–Douglas production technology. Gagnon and Kahn (2005) consider other production structures where marginal cost is not measured with labour's share and conclude that the Galí and Gertler (1999) findings hold up. Galí and Gertler estimate of θ implies more than a year between price changes, but they cannot reject the equation. Micro price data might be useful to identify θ, but, under the frequency of *re-optimization* interpretation of the Calvo model, estimates of θ over a year for the United States seem too high (cf. Blinder et al. 1998). Galí and Gertler consider an alternative model with 'rule-of-thumb' firms who use lagged inflation to update their prices when they have the opportunity to re-optimize. This model is motivated by the fact that lagged inflation enters significantly in the empirical version of (1). The model with rule-of-thumb firms is not rejected and the estimates of θ imply prices are re-optimized every two or three quarters. The latter estimates are within the range of plausibility. Galí and Gertler also estimate the number of rule-of-thumb firms to be small and emphasize that (1) holds approximately. Bakhshi et al. (2005) argue that the Galí–Gertler 'hybrid' model is a good approximation to Dotsey et al.'s (1999) menu cost model.

It is clear that θ is not identified separately from A and D in (1). However, A and D can be identified with auxiliary information. Sbordone (2002) identifies D by assuming the stock of capital is fixed exogenously at each firm for all time. Under the usual assumption of a Cobb–Douglas production function, auxiliary information on the share of labour income in GDP can be used to identify D. Sbordone considers the forward looking solution to (1) as well as the solution to a similar equation for the labour market. The expected present-value calculations needed to implement this estimation are implemented with a vector autoregression. This empirical strategy is analogous to Abel and Blanchard's (1986) approach to estimating investment adjustment costs. Sbordone estimates prices are re-optimized every one to two quarters. Galí et al. (2001) apply Sbordone's fixed capital assumption to their rule-of-thumb model and estimate the frequency of re-optimization to be a little higher than in Galí and Gertler (1999), and significant small positive numbers of rule-of-thumb firms continue to be estimated.

Eichenbaum and Fisher (2007) explore Woodford's (2003, 2005) dynamic version of Sbordone's (2002) model in an environment which also includes Kimball's (1995) variable markup. As in Galí and Gertler (1999) and Galí, Gertler and López-Salido. (Galí et al. 2001), they adopt a GMM estimation and testing strategy. To improve the power of their tests, Eichenbaum and Fisher (2004) impose the restrictions Eq. (1) place on the moving average structure of the Euler equation errors and reduce the number of instruments compared to the previous papers. They easily reject (1) assuming the Euler error is an MA(0). This motivates them to include the auxiliary assumption that firms make decisions based on lagged information. With one such *implementation* lag, this yields an MA(1) structure which is not rejected and which the re-optimization frequency is about two years, if $A = D = 1$. With empirically motivated values for the curvature of the demand curve and the size of capital adjustment costs, re-optimization every two quarters cannot be ruled out at conventional significance levels. Eichenbaum and Fisher include dynamic indexation (prices indexed to the most recent inflation rate) of the prices of firms that do not re-optimize in a given period. This is an alternative to rule-of-thumb firms as away of introducing a lagged inflation term into (1). Eichenbaum and Fisher (2004) find they cannot reject the possibility that there are no rule-of-thumb firms under dynamic indexation.

Recently much work has been done to document prices at the microlevel. Blinder et al. (1998) survey actual price setters and find that, among firms reporting regular price reviews, annual reviews are by far the most common. Other key contributions include Bils and Klenow (2004), Klenow and Krystov (2005) and work done with European data for example by Stahl (2005). Much of this literature emphasizes the frequency of price changes. For example, Blinder et al. (1998) report that the median time between price changes among the firms that they survey is roughly three quarters. Comparing the Calvo–Yun Phillips curve with these findings is delicate. With price indexation the model implies that prices change *too* frequently relative to the micro data because all prices are changing all the time. Also, just because firms are changing prices does not mean that they have re-optimized those prices: a subset of the price changes being recorded could reflect various forms of time-dependent pricing rules.

Integrating over all the micro evidence, with a low inflation economy like the United States, versions of the Calvo–Yun Phillips curve with an implementation lag, dynamic indexation, capital adjustment costs and time-varying markups can be reconciled with the macro data without requiring implausible degrees of rigidities in price-setting behaviour at the micro level. Of course this model is not literally ‘true’. For instance, the model also has the implausible implication that any CPI observation for which $P_{i,t}/P_{i,t-1}$ is not equal to π_{t-1} involves re-optimization. Developing tractable models that are fully consistent with the salient macro facts and the emerging literature on the behaviour of individual good prices is a key challenge going forward.

See Also

- ▶ Adaptive Expectations
- ▶ Adjustment Costs
- ▶ Generalized Method of Moments Estimation
- ▶ Inflation
- ▶ Inflation Dynamics
- ▶ Macroeconomic Forecasting
- ▶ Phillips Curve

Bibliography

Abel, A., and O. Blanchard. 1986. The present value of profits and cyclical movements in investment. *Econometrica* 54: 249–274.

Atkeson, A., and L.E. Ohanian. 2001. Are Phillips curves useful for forecasting inflation? *Federal Reserve Bank of Minneapolis Quarterly Review* 25(1): 2–11.

Bakhshi, H., Kahn, H. and Rudolf, B. 2005. *The Phillips curve under state-dependent pricing*. Manuscript, Carleton University.

Bils, M., and P. Klenow. 2004. Some evidence on the importance of sticky prices. *Journal of Political Economy* 112: 947–985.

Blinder, A., E. Canetti, D. Lebow, and J. Rudd. 1998. *Asking about prices: A new approach to understanding price stickiness*. New York: Russell Sage Foundation.

Brave, S., and J.D.M. Fisher. 2004. In search of a robust inflation forecast. *Economic Perspectives* 28(4): 12–31.

Calvo, G. 1983. Staggered prices in a utility-maximizing framework. *Journal of Monetary Economics* 12: 383–398.

Christiano, L., and T. Fitzgerald. 2003. The band pass filter. *International Journal of Economics* 44: 435–465.

Christiano, L., M. Eichenbaum, and C. Evans. 2005. Nominal rigidities and the dynamic effects of a shock to monetary policy. *Journal of Political Economy* 113: 1–45.

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

Eichenbaum, M. and Fisher, J. 2004. *Evaluating the Calvo model of sticky prices*. Working Paper No. 10617. Cambridge, MA: NBER.

Eichenbaum, M., and J. Fisher. 2007. Estimating the frequency of price re-optimization in Calvo-style models. *Journal of Monetary Economics* 54(7): 2032–2047 (forthcoming).

Erceg, C., J. Henderson, W. Dale, and A.T. Levin. 2000. Optimal monetary policy with staggered wage and price contracts. *Journal of Monetary Economics* 46: 281–313.

Fisher, J.D.M. 2006. The dynamic effects of neutral and investment-specific technology shocks. *Journal of Political Economy* 114: 413–451.

Fisher, J.D.M., C. Liu, and R. Zhou. 2002. When can we forecast inflation. *Economic Perspectives* 26(1): 30–42.

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

Gagnon, E., and H. Kahn. 2005. New Phillips curve under alternative production technologies for Canada, the United States, and the Euro area. *European Economic Review* 49: 1571–1602.

Galí, J., and M. Gertler. 1999. Inflation dynamics: A structural econometric analysis. *Journal of Monetary Economics* 44: 195–222.

Galí, J., M. Gertler, and D. López-Salido. 2001. European inflation dynamics. *European Economic Review* 45: 1237–1270.

Galí, J., M. Gertler, and D. López-Salido. 2007. Mark-ups, gaps and the welfare costs of business cycles. *Review of Economics and Statistics* 89: 44–59.

Gertler, M., and J. Leahy. 2005. *A Phillips curve with an Ss foundation*. New York: University manuscript.

Golosov, M. and Lucas, R.E., Jr. 2003. *Menu costs and Phillips curves. Working Paper No. 101187*. Cambridge, MA: NBER.

Hansen, L.P. 1982. Large sample properties of generalized method of moments estimators. *Econometrica* 50: 1029–1054.

Kimball, M. 1995. The quantitative analytics of the basic neomonetarist model. *Journal of Money, Credit, and Banking* 27: 1241–1277.

Klenow, P. and Krystov, O. 2005. *State-dependent or time-dependent pricing: Does it matter for recent US inflation?* Working paper, Bank of Canada.

Lucas, R.E. Jr. 1972. Expectations and the neutrality of money. *Journal of Economic Theory* 4: 103–124.

Lucas, R.E. Jr. 1978. Asset prices in an exchange economy. *Econometrica* 46: 1429–1445.

Midrigan, V. 2005. *Menu costs, multiproduct firms and aggregate fluctuations*. Manuscript, Ohio State University.

Phillips, A.W. 1958. The relation between unemployment and the rate of change of money wage rates in the United Kingdom, 1861–1957. *Economica* 25: 283–299.

Samuelson, P.A., and R.M. Solow. 1960. Analytical aspects of anti-inflation policy. *American Economic Review* 50: 177–194.

Sbordone, A. 2002. Prices and unit labor costs: A new test of price stickiness. *Journal of Monetary Economics* 49: 265–292.

Stahl, H. 2005. *Price setting in German manufacturing: New evidence from new survey data*. Working Paper No. 561, European Central Bank.

Stock, J., and M. Watson. 1999. Forecasting inflation. *Journal of Monetary Economics* 44: 293–335.

Taylor, J. 1980. Aggregate dynamics and staggered contracts. *Journal of Political Economy* 88: 1–23.

Woodford, M. 2003. *Interest and prices: Foundations of a theory of monetary policy.* Princeton: Princeton University Press.

Woodford, M. 2005. Firm-specific capital and the New Keynesian Phillips curve. *International Journal of Central Banking* 1(2): 1–46.

Yun, T. 1996. Nominal price rigidity, money supply endogeneity, and business cycles. *Journal of Monetary Economics* 37: 345–370.

Phillips, Alban William Housego (1914–1975)

C. A. Blyth

Keywords

Identification; Inflation; Phillips curve; Phillips, A. W. H.; Stabilization policy; Unemployment

JEL Classifications

B31

Bill Phillips was born on 18 November 1914 into a farming family in Te Rehunga, near Dannevirke, in southern Hawkes Bay in the North Island of New Zealand, and died at Auckland, New Zealand on 4 March 1975. He came to economics after a career as an electrical engineer and following military service and imprisonment by the Japanese in the Second World War; in 1946 he became a Member of the Order of the British Empire for his military services. His rise in the profession was rapid. He was appointed an Assistant Lecturer at the London School of Economics in 1950 and to a Readership in 1954. In 1958 he became Tooke Professor, resigning in 1967 to take a Chair at the Institute of Advanced Studies, Australian National University in Canberra. A crippling stroke in 1969 forced his retirement and he lived in Auckland until his death. In his short career in economics he made major contributions to problems of dynamic stabilization, estimation and, most notoriously, empirical economics, where he gave his name to the ‘Phillips curve’.

Phillips’s Ph.D. at LSE was on the problems of stabilizing or controlling an economy. Before this he had built a hydraulic model in perspex of a dynamic Keynesian-type economy and sold commercial versions of the model to academic and other institutions in Britain and the United States. (Some machines had a bottle of water named the Bank of England, used for ‘topping up’. Richard

Goodwin is credited with introducing an accelerator in Cambridge's machine, but leakages were always a problem.) In a seminal paper in 1954 Phillips dealt with response lags and the problems they presented for stabilization policy. He distinguished between proportional, integral and derivative policies, depending on whether policy changes responded to current errors, cumulated deviations or rates of change of objectives. Optimal policy depends on the lag properties of the economy, and would consist of a mixture of proportional, integral and derivative components. Subsequent analysis of stabilization policy has used this scheme (for example, Meade 1971).

This early work convinced Phillips that proper econometric modelling was a precondition for dynamic stabilization. He turned to the problem of empirical description of the lag structures and their statistical estimation. All his later papers are concerned with the problems and difficulties of describing and estimating the dynamic relationships embodied in time forms of economic responses. His retreat to Canberra and Chinese economic studies has been seen by Lancaster (1979) as an acknowledgement that he found the problems, as he posed them, of estimating the relationships required for dynamic control beyond his capacity to solve. In Canberra, however, he continued to work on problems of identification (for example, 1968). In this later work he foreshadowed subsequent thinking (for example, that of Lucas) by explaining how the application of stabilization policy through a model results in the model becoming underidentified or, more generally, how policy changes relationships. The 'Phillips dilemma' remains: in the absence of adequate econometric modelling, stabilization policy is an empty box.

But before he left macroeconomics Phillips made in 1958 his epochal contribution of the Phillips curve, which mesmerized economists for the next decade and continues to attract attention. Responding to Dennis Robertson's criticism of the Keynesian mathematical model of his Ph.D. thesis, Phillips later used a relation between the rate of price change and capacity utilization, but without being able to give it any satisfactory empirical foundation. The lengthy time series of British wages produced by Henry Phelps Brown and Sheila Hopkins gave him the opportunity to experiment with the long series of British unemployment statistics from 1861 to 1957. What began as an attempt to derive a simple relationship between rate of change of wage rates and unemployment emerged as a nonlinear long term relationship with a complex short period lagged response. The famous paper is striking for the informal estimation method and the ad hoc theorizing and Phillips admitted an excessive haste to publish while also acknowledging that A.J. Brown had almost got the results earlier but without the lags (Blyth 1975). Apart from an enquiry into Australian statistics, Phillips did not enter into the subsequent international controversy over the theoretical and empirical foundations of the 'Phillips curve'. It is necessary to read Phillips's original paper to understand its intentional exploratory character, and the deliberate absence of theoretical generalization. For an enquiry which in the eyes of, for example, Samuelson and Solow 'closed' the Keynesian system, it is remarkably but typically modest and tentative. Furthermore, on the controversial issue of the high cost of the trade-off between inflation and unemployment, Phillips refers briefly to the 5 per cent unemployment level necessary to maintain stable wage rates without expanding on it. If there were close intellectual connections with the Joan Robinson–Kalecki–Beveridge approach to the full employment–trade union problem, they are not disclosed.

Scientific problems with stabilization theory, family reasons and reaction to student revolt at LSE all may have contributed to Phillips's move to Australia, where he energetically began to develop Chinese economic studies. His interest in China had begun in the 1930s and he began to learn Chinese while a prisoner-of-war in Java. In the short period before he retired he saw the firm establishment of a Centre for Contemporary Chinese Studies in Canberra, while his final academic activity in Auckland was appropriately enough to start a lecture course in Chinese economic history. In 1974, on his 60th birthday, his colleagues and friends presented him with a subsequently published Festschrift (Bergstrom et al. 1978).

In a profession which accepts many strays from other disciplines, Phillips was outstanding. His formal education ended in New Zealand at the age of 15, and after apprenticeship as an electrician he qualified as an electrical engineer in London in 1938 after working in Australia and travelling to Britain via Japan and Siberia. An authentic hero of the Second World War, his introduction to economics was as he said through a poor degree in sociology. James Meade sponsored his hydraulic model, and Phillips's 'launching' is by tradition associated with a famous Robbins seminar in which he successfully explained his machine. A New Zealand influence and connection is not intellectually evident, but may have contributed to the willingness to 'do it himself'. Who else but a New Zealander would have learnt his differential equations at a Queensland goldmine?

See Also

- ▶ Phillips Curve
- ▶ Phillips Curve (New Views)

Selected Works

1950. Mechanical models in economic dynamics. *Economica* 17: 283–305.

1954. Stabilization policy in a closed economy. *Economic Journal* 64: 290–323.

1956. Some notes on the estimation of time-forms of reactions in interdependent dynamic systems. *Economica* 23: 99–113.

1957. Stabilization policy and the time-forms of lagged responses. *Economic Journal* 67: 265–277.

1958. The relation between unemployment and the rate of change of money wage rates in the United Kingdom, 1861–1957. *Economica* 25: 283–299.

1959. The estimation of parameters in systems of stochastic differential equations. *Biometrika* 46: 67–76.

1968. Models for the control of economic fluctuations. In Scientific Control Systems, *Mathematical model building in economics and industry.* London: Griffin.

Bibliography

Bergstom, A.R., Catt, A.J.L., Peston, M.H., and Silverstone, B.D.J. (eds.). 1978. *Stability and inflation.* Chichester: Wiley.

Blyth, C.A. 1975. A.W.H. Phillips. *Economic Record* 51: 303–307.

Lancaster, K. 1979. A. William Phillips. In *International encyclopaedia of the social sciences*, vol. 18. New York: Free Press.

Meade, J.A. 1971. *The controlled economy.* London: Allen & Unwin.

Philosophy and Economics

D. Wade Hands

Abstract

The literature on philosophy and economics has traditionally been divided into two areas: economic methodology, which connects economics and epistemology/philosophy of science, and the literature on economics and moral philosophy/ethics. Recent developments in both of these areas are discussed in detail.

Keywords

Altruism; Bargaining; Capabilities approach to social welfare; Consumer choice theory; Critical realist research programme; Economics of scientific knowledge; Empirical macroeconomics; Epistemology and economics; Ethics and economics; Evolutionary biology; Experienced utility; Experimental economics; Falsificationism; Free-rider problem; Happiness; Hedonism; History of economic thought; Human Development Index; Hume, D.; Hutchison, T.; Interpersonal utility comparisons; Justice; Kahneman, D.; Material welfare school; Methodology of economics; Models; Naturalism; Neuroeconomics; Ontology and economics; Pareto efficiency; Philosophy and economics; Philosophy of science; Popper, K.; Positive–normative dichotomy; Positivism; Postmodernism; Preferences; Rational choice theory; Rawls, J.; Robbins, L.; Sen, A.; Social

contract; Ultimatum game; Utilitarianism; Value judgements; Welfare economics

JEL Classifications

B1

The essential interdependency of philosophical and economic ideas was a prominent feature of classical economics. Adam Smith was the author of *The Theory of Moral Sentiments* as well as *Wealth of Nations.* John Stuart Mill was an extremely wide-ranging scholar, as well known as the author of *A System of Logic* as of *The Principles of Political Economy.* And of course Karl Marx's *Capital* also drew on intellectual resources from economics, philosophy and a number of other fields. Classical political economy was deeply influenced by philosophy – different philosophies for different economists, but influenced nonetheless – and ideas also flowed freely in the opposite direction, from political economy to various areas of philosophical inquiry.

This changed significantly in the first third of the 20th century. The abandonment of 'political economy' and the self-conscious development of 'scientific economics' coincided with a major change in the relationship between the two disciplines. Although philosophy never completely disappeared from economic theorizing, it systematically came to play a less and less obvious role. There are undoubtedly many reasons for this. Two of the more important include the overall professionalization of disciplinary economics and the general acceptance of a more narrow, positivist-inspired notion of legitimate 'scientific' inquiry. John Stuart Mill directed his arguments at the general educated public and wrote confidently about the 'moral sciences'; by the first half of the 20th century fewer economists were doing the former and almost no professional economist would be comfortable doing the latter.

Although there were different versions of positivism, one common theme was that 'meaningful' discourse comes in only two forms: the synthetic knowledge of empirical science and the analytic knowledge of logic and mathematics. During the period of positivist dominance (roughly from the early 1930s through the 1950s), many, perhaps most, of the lines of inquiry that had previously travelled under the label of 'philosophy' – including, ethics, ontology, metaphysics, and aesthetics – were dismissed from the realm of meaningful discourse. Science ceased to be a generic category that included any rational, non-faith-based inquiry, and instead came to designate only the natural sciences (or modes of inquiry that follow the same scientific method). Economics clearly had scientific aspirations, and in such a regime fulfilling those aspirations required jettisoning the profession's old philosophical ways. Many of the significant developments in economic theory during the first half of the 20th century can be understood in precisely these terms: as an attempt to systematically discard the old metaphysical and utilitarian baggage, and replace it with more appropriate scientific concepts. Moral philosophy, for example, might still make an appearance in discussions about economic theory, but it almost always played a disparaging role: either to indict another theory for retaining some ethical residuum, or to emphasize that one's own theory was entirely free of such normative influences. Such an environment was certainly not conducive to forging new links between philosophy and economics, and for much of the 20th century very few were.

A particularly good example of the rejection of philosophy is the development of welfare economics during the second quarter of the 20th century. From the hedonism of many early neoclassicals to the so-called 'material welfare school' (Cooter and Rappoport 1984) of Alfred Marshall and Arthur Pigou, welfare economics (and applied microeconomics in general) had traditionally been associated with utilitarianism: policy A was better than policy B if A increased total utility by more than B. During the 1930s, as a result of the work of Lionel Robbins (1952) and others, most economists came to view this type of 'interpersonal' utility comparison as unscientific and thus inappropriate for economic analysis. Moral values were simply raw, subjective or 'emotive' preferences that were not amenable to scientific analysis, and must therefore be kept out of economic science.

As economists moved away from the earlier utilitarian notions of 'good' economic policy, they increasingly turned to the Pareto criterion as an alternative evaluative standard. It was argued (and still is) that Pareto efficiency – an allocation of resources such that no one person can be made better off without making someone else worse off – does not require making interpersonal utility comparisons and is therefore an entirely appropriate standard for scientific economics. The most important theoretical results of modern welfare economics – the first and second fundamental theorems – are based on a direct application of the Pareto criterion to questions about the welfare implications of competitive equilibrium. Although the norm-free credentials of Pareto efficiency have repeatedly been challenged (Blaug 1980; Hausman and McPherson 2006; Robertson 1952), the standard interpretation among practising economists remains that such judgements, and thus any policy recommendations based on them, are fundamentally value free. But it is not necessary to take sides in the debate over whether Pareto efficiency is or is not an ethical criterion in order to recognize that the entire discussion is couched in terms of whether moral concepts are properly kept out of economic science, and to note that such a discussion does not provide a very fertile environment for the cultivation of new relationships between economics and moral philosophy.

Economic methodology has traditionally been the one exception to economists' general rejection of philosophy. Although ethics and metaphysics were shunned by economists, epistemology and philosophy of science were often consulted for guidance regarding the proper scientific method. This said, even within methodology the use of philosophical resources varied greatly from economist to economist. Some of the classical works in economic methodology (Milton Friedman 1953, for example) hardly mentioned philosophy at all; others (Robbins 1952, and Hutchison 1938, for example) drew on selected aspects of the philosophy of science, while still others (Blaug 1980; Samuelson 1963) tried to apply the arguments of particular philosophers of natural science directly to economics. Thus, even in methodology economists focused on only a relatively small portion of the philosophical literature and employed even those resources in a less than systematic way.

Although the traditional methodological literature is both extensive and ongoing, it is not the focus of the following discussion. There are at least two reasons for this. First, this literature has been effectively surveyed in a number of contemporary works (Blaug 1980; Caldwell 1994; Hands 2001; Hausman 1992) and second, things have again changed. Since the mid-1980s there has been a renaissance in the interaction between economics and philosophy. The traditional approach to economic methodology continues to produce viable research, but economics and philosophy are also interacting in many other, new and important ways. Philosophy of natural science is no longer the only relevant set of philosophical ideas – ethics and ontology have both returned to the scene – and the intellectual dynamic is now one of bilateral exchange rather than economists simply borrowing ideas from one corner of the philosophical shelf.

In addition to the revival of the interplay between economics and philosophy there has been an increase in the traffic between economics and a number of other fields that compete for some of the same intellectual space that philosophy has traditionally occupied. For example, resources from the sociology of science and science studies (Mirowski 2002; Sent 1998; Weintraub 2002; Yonay 1998), the rhetoric of science (McCloskey 1998), postmodernism (Ruccio and Amariglio 2003), feminism (Ferber and Nelson 2003; Nelson 1996), and variety of other fields have provided new tools for the examination of (and often confrontation with) modern economic theory. Although these works frequently overlap with the literature on philosophy and economics, they also involve ideas sufficiently removed from disciplinary philosophy that they fall outside of the work considered here.

The discussion is divided into two parts; the first examines recent developments in the relationship between economics and scientific philosophy. Some of this work has much in common with traditional economic methodology, while other contributions approach the relationship in entirely new ways. In the interest of brevity, only five of the many possible areas of significant research are examined. The second section examines the recent literature that combines economics and moral philosophy. Ethical questions are again back on the table, and an extensive literature has grown up relating various issues in moral philosophy to developments within economic theory. Some of this research challenges the received view of the relationship between economics and ethics established during the first half of the 20th century, while other parts of the literature develop totally new connections. Again, as with the methodological literature, only a few examples are discussed. The final section briefly considers some points of convergence between contemporary work on economics and epistemology and that on economics and ethics. Throughout the discussion, the emphasis is on microeconomics and rational choice theory (rather than, say, macroeconomics or econometrics).

Economics, Epistemology, and Philosophy of Science

The first area of research to be examined goes back to Terence Hutchison (1938); it is the literature relating the philosophical ideas of *Karl Popper* (1965, 1968) to economics. Popper is best known as an advocate of falsificationism, a philosophy that has two main theses: one demarcating science from non-science and the other characterizing the growth of scientific knowledge. For a theory to be scientific it must be at least potentially falsifiable by empirical evidence (in Popperian language, be falsifiable by at least one empirical basic statement). Scientific knowledge grows as the scientific community rejects falsified theories and retains those that have survived attempted falsifications (that is, by 'bold conjecture and severe test'). The body of accepted science at any point in time consists of all scientific theories that have survived such severe empirical tests. Elements of such a methodology were present in Hutchison (1938), and elaborated in more detail in his later work. The position has been most articulately defended in the methodological writings of Mark Blaug (1980). Although many economists continue to endorse a falsificationist approach to methodological questions, there is also an extensive critical literature on the subject (Caldwell 1991, 1994; Hands 1993; Hausman 1988, 1992).

If the only research connecting the Popperian tradition to economics was the literature on falsificationism, then the subject would probably not be included in this discussion of recent developments. But that is not the case. During the last few decades the Popperian tradition has engaged economics on a number of different fronts, and currently consists of much more than just the literature defending (or criticizing) falsificationism (Caldwell 1991). At least three other developments should be noted. The first involves Popper's own brief discussion of economic methodology (Popper 1994). This work is controversial because Popper's statements about economics – and social science more generally – differ from what he said about the (falsificationist) methodology of natural science. The second concerns the so-called 'critical rationalist' interpretation of Popper's overall philosophical programme: an interpretation that goes back in the economics literature to Kurt Klappholz and Joseph Agassi (1959), but has its best contemporary representation in the work of Lawrence Boland (1997). Supporters of critical rationalism argue that Popper's main philosophical contribution was not (empirical) falsificationism but rather a more general view of the growth of knowledge through open debate and rational criticism – of which falsification by empirical evidence is simply one, albeit a very important, special case. Although the discussion of critical rationalism has remained primarily an in-house debate among Popperians, it has much broader implications because it opens the door to characterizing the growth of knowledge as a product of particular social institutions rather than as the result of following fixed methodological rules, a view that has become increasingly important in general philosophy of science. Finally, there has been an extensive discussion of the work of Popper's student Imre Lakatos (1970) and his 'methodology of scientific research programs' (Backhouse 1997; Blaug and De Marchi 1991; Latsis 1976). Economists have focused on two different aspects of Lakatos's work: his historical framework for understanding the evolution of economic research programmes (his concepts of hard core, protective belt, and so on) and his specific methodological framework for appraising scientific research programmes as progressive or degenerating. Even though there exists a critical literature on both of these issues, the Lakatosian framework has produced important case studies and also encouraged a re-examination of the general relationship between economic methodology and the history of economic thought.

The second area to consider involves the revival of interest in *ontology* and *metaphysics* in the philosophy of economics. There now exists a burgeoning literature on 'economics and ontology' (Mäki 2001), something that would have been next-to-impossible only a few decades ago. During the heyday of positivism any mention of such (occult) notions as essential natures, underlying causal powers, or ontological necessity all but disappeared from academic discussions about economics. Ontological discussion continued to some extent within certain heterodox, particularly Marxist, research programmes, but among mainstream economists, even philosophically informed ones, such concepts had no place in professional discourse. Although many things have contributed to this revival, three issues seem to be particularly important.

One factor contributing to this ontological renewal has clearly been the development of the 'critical realist' research programme, an anti-empiricist approach to the philosophy of social science that focuses on uncovering the hidden underlying causal mechanisms at work in social life. The most prolific defender of critical realism within economics has been Tony Lawson (2003), and his writings have generated an extensive secondary literature. A second factor involves changes that have taken place within the philosophy of natural science. Although there were many reasons for the decline of positivist-inspired philosophy of science, one of the most important was the perception that serious problems had developed within the Humean-inspired 'empiricist' component of the programme. Although debate continues about whether the founders of positivism were actually as empiricist as the standard view suggests (Michael Friedman 1999), it is certainly clear that the programme was perceived that way by both critics and supporters, and that it was this aspect of the programme that was most effectively targeted by the criticism that descended upon it in the last quarter of the 20th century. Some of the efforts to reconfigure our reigning philosophical conceptions in light of these developments – particularly about scientific laws (Cartwright 1989) and causality (Hoover 2001) – draw directly on insights from economics. Finally, the literature on economics and ontology has benefited from recent changes that have taken place within the discipline of economics itself. A discipline that is more willing to entertain theoretical pluralism is more likely to be willing to entertain philosophical, even ontological, pluralism as well. The bottom line is that ontology and metaphysics are back and they are opening up a number of new (and renewed) lines of inquiry relevant to the philosophy of economics.

The third set of changes to consider involves border crossings between economics and certain other scientific fields – *cognitive science, neuroscience,* and related disciplines – that have influenced the recent literature on the philosophy of mind. This literature is relatively new and rapidly growing, so much so that no appellative convention has emerged. Until such a consensus has been reached it is perhaps best to be inclusive and simply call it the literature on 'the mind, the brain, rationality, agency and economics'. Examples would include such disparate works as Davis (2003), Glimcher (2003), Mirowski (2002), and Ross (2005). Although the arguments of the various contributors are quite different, there is some agreement about the main issues, as well as about the requirements for any adequate approach to these issues. These requirements concern consistency with recent developments in fields such as cognitive science, neurophysiology and artificial intelligence. The common concern is the core rational choice framework of modern economics: explaining economic behaviour as the outcome of rational constrained optimization of well-ordered preferences. Consumer choice theory is the paradigm case of such an explanatory strategy, but it is standard throughout economics (traditionally microeconomics, but increasingly macroeconomics as well).

Such rational choice explanations have recently been subject to a variety of criticisms. Some of these relate to the abundance of contrary empirical evidence that has appeared in the experimental literature – in both economics and psychology (Kahneman and Tversky 2000) – and some of it has to do with the well-known philosophical problems associated with 'intentional' or 'folk psychological' (belief-desire-action) explanations (Rosenberg 1992). Although much of the impetus comes from critiques of rational choice theory, this does not mean that all of the resulting literature advocates doing away with it. Some authors clearly do, but others interpret these recent theoretical developments as a way of defending standard practice. In either case, whether its authors defend or attack rational choice theory, the literature embodies a fundamental change in the rules of engagement. It is too early to know how it will develop, or the various turns it might take along the way, but it is clear that both in its use of resources from other disciplines and in its overall mode of argumentation it has moved economics and philosophy in a substantially different direction.

The fourth area to consider overlaps substantially with previous section on minds, brains, cognitive science, and such. It concerns the tendency towards *'naturalism'* in epistemology and philosophy of science. The standard interpretation of both positivist and falsificationist philosophy of science puts 'philosophy before science' in the sense that philosophers first decide what scientists must do to produce theories that are cognitively significant – constitute legitimate scientific 'knowledge' – and then evaluate specific scientific practices on the basis of this philosophical analysis. Naturalism – and there are many specific versions, but here we consider its most generic form – reverses this relationship. Instead of starting with a priori philosophical analysis about what scientific knowledge must be, naturalism starts with science, that is, the best current scientific practice, and uses this best practice to inform our epistemological inquiries about knowledge in general. Much of the philosophical literature discussed in the previous section – the literature that employs contemporary cognitive science and neuroscience in the investigation of knowledge in general – is naturalist in this sense. Such naturalism raises a host of questions, particularly questions about how it is possible to have a 'normative' philosophy of science, one that explains what ought to be done in science, when the 'philosophy' in question is based on descriptions of scientific practice. Such questions are the subject of much current debate and do not have easy or simple answers. Fortunately, such answers are not required for a discussion of how naturalism has affected research in the philosophy of economics.

Much of the recent research in the history and philosophy of economics is broadly naturalist in spirit. Naturalism informs some of the work on traditional methodological questions (Hausman 1992) as well as research in general philosophy of science that draws heavily on economics (Cartwright 1989). It also provides the backdrop for a number of recent studies on specific research programmes within economics, including the role of models (Morgan 1999, 2001), the practice of empirical macroeconomics (Hoover 2001), and the development of experimental economics (Guala 2005). Although the boundary that separates such naturalist-inspired research from similar work informed by science studies is somewhat blurred, it is often possible to categorize a particular piece of work as primarily one or the other. If the main question is the philosophical justification of the particular economic tool or theory – even if the standards for such justification are naturalistically or historically grounded – then the research is in the spirit of naturalistic philosophy; but if the explanation of the acceptance or rejection of particular economic tools or theories is based primarily on the influence of social, political, or individual (non-epistemic) interests, then it falls more into science studies.

The final category of literature to be considered, the *economics of scientific knowledge,* reverses the standard relationship between a particular social science like economics and the philosophy of natural science. As discussed above, the traditional relationship between philosophy of science and economics has been that philosophy comes first (laying the foundations for knowledge), economic methodology then translates those philosophical ideas into the context of economic science, and finally particular economic theories are appraised on the basis of the methodological rules so acquired. In the economics of scientific knowledge this process is reversed. Certain areas of economic theory – for example, industrial organization (IO) economics – examine how the institutional organization of a particular industry contributes to economic efficiency. Shifting this type of reasoning from the production of goods and services to the production of scientific knowledge is the basis for one way of thinking about the economics of scientific knowledge. The scientific community has a particular institutional structure; if the goal of this scientific 'industry' is the production of (reliable, justified, . . .) scientific knowledge, then an obvious question is the degree to which the industrial organization contributes to the growth of knowledge (that is, epistemic efficiency). Since the goal is the growth of knowledge within the community, it might be the case that all of the individual scientists following the same methodological rule may not be the optimal way to arrange the available epistemic resources; perhaps the greatest production of scientific knowledge comes about as the result of a 'cognitive division of labor' (Kitcher 1993) rather than methodological homogeneity. It is easy to see how such an approach opens up new ways of thinking about the growth of scientific knowledge, and does so by employing economic theory as a resource (in the spirit of naturalism) to address general questions about the growth of knowledge and the optimal design of scientific institutions.

It can be argued that such research on the economics of scientific knowledge goes back to Charles Sanders Peirce in 1879 (Wible 1998), but regardless of its origins it has expanded rapidly during the last few years, with contributions coming from both economists and philosophers (Dasgupta and David 1994; Goldman and Shaked 1991; Kitcher 1993; Wible 1998). As one might expect, the literature has also generated a variety of critical responses (Hands 1997; Mirowski 2004). In addition, many other contributions to the economics of scientific knowledge are quite different from the version of epistemic IO discussed above (Mirowski and Sent 2002). But in all of its various forms this work clearly represents a significant change in the interaction between economics and philosophy of science.

Economics and Moral Philosophy

One of the many changes that have taken place in the relationship between economics and moral philosophy has been a re-examination of economists' traditional stance on the 'positive-normative dichotomy'. This change is sufficiently complex that it is examined in two parts. First, there has been a substantive reconsideration of the general place of 'the normative' within the science of economics (where 'normative' does not necessarily concern ethics), and second, ethical norms are increasingly being considered in the causal explanation of economic phenomena.

Enforcing the prohibition against value judgements in economics requires maintaining a strict dichotomy between positive statements about what 'is' and normative statements about what 'ought to be'. These two issues – dichotomization and prohibition – are certainly related, but they can also be separated. The first asserts that a dichotomy should be maintained – 'ought' should be kept separate (and cannot be derived) from 'is' – while the second asserts that separate is not equal – things on the normative/'ought' side of the dichotomy have no place within scientific economics. Although the first (dichotomy) is necessary for the second (prohibition), it is clearly not sufficient; one could argue, as, say, Mill and Marshall did, that there is a difference between positive and normative economics, and yet also leave room for a version of normative economic science.

Debate over the strict dichotomy and the prohibition against deriving 'ought' from 'is' has a long history. It was popularized by David Hume in the 18th century, labelled the 'naturalistic fallacy' by G.E. Moore early in the 20th century, and is the subject of a long and contentious debate within philosophy (Putnam 2002). Although many economists have been concerned with these issues, the one who probably played the most important role in the profession's ultimate establishment of the principle of strict separation was Lionel Robbins. Robbins (1952, p. 149) endorsed a strict dichotomy – 'Propositions involving the verb "ought" are different in kind from propositions involving the verb "is"' – but he went beyond mere separation to prohibition, advocating complete exclusion of normative analysis from scientific economics. In particular, he criticized the normative welfare economics of the Marshallian school because it relied on 'interpersonal' utility comparisons. For Robbins, the normative economics resulting from such analysis was 'illegitimate' and 'lacking in scientific foundation' (1952, p. 141).

By and large Robbins's position on these matters has become the conventional wisdom among practising economists as well as among most contributors to the methodological literature. Where methodological commentators often differ is not over whether normative concerns should be kept out of scientific economics but rather on the factual question of whether most practising economists have actually done so. For example, two well-known contributors to economic methodology, Mark Blaug (1980) and Milton Friedman (1953), both endorse the dichotomy and prohibition, but differ on the question of whether the economics profession has in fact been successful at keeping normative propositions out of its scientific practice.

The core of standard microeconomics continues to be rational choice theory; economic agents are assumed to have well-ordered preferences and make optimal choices given those preferences and the various constraints they face. Such rational choice explanations involve two parts: preferences (goals/ends) are assumed to be rational (that is, well-ordered, satisfying conditions such as transitivity and completeness) and the agent is presumed to act in the most efficient way to achieve those given ends (that is, to act in an instrumentally rational way). Philosophers have traditionally called such rationality 'practical rationality' to distinguish it from 'theoretical' or 'epistemic' rationality. In general practical rationality involves what it is rational to do, or at least intend to do, while theoretical or epistemic rationality involves what it is rational to believe.

The literature on practical rationality leads to a very different characterization of the positive–normative dichotomy than the one standard in economics. Although most practising economists continue to view rational choice theory as a positive theory about the behaviour of economic agents (at least under ideal conditions), most philosophers writing on the subject consider it a normative theory in the sense that it involves norms and obligations. Practical rationality, and thus rational choice theory as a particular instantiation of it, is a normative theory because it tells agents what they 'ought' to do in order to act rationally. In the contemporary philosophical literature this view is often associated with the work of Donald Davidson (2001), but it has a long history and continues to be debated (Searle 2001). Philosophers have certainly not closed the book on the question of how a theory of practical rationality could be a descriptive theory, or how, if it is normative, it might relate to associated descriptive theories. The point is simply that it is increasingly the case, in both philosophy and economics, that the discussion of rational choice theory starts from the presumption that it is a particular instantiation of the theory of normative rationality, and as a result, the description of actual economic agents – whether in the laboratory or in 'the wild' – is coming to be seen as something to be compared with, or reconciled with, this theory of normative rationality. It is still possible to discuss the ways in which rational choice theory is or is not an adequate scientific theory, but the starting point of the discussion has changed substantially (Hausman and McPherson 2006; Mongin 2006; Ross 2005).

The second change to be examined requires us to step back from the previous discussion of normative rationality. Suppose we use 'normative' to mean 'ethically normative', and view rational choice theory as a strictly positive, not a normative, theory, then *there are still a number of arguments for increasing the normative content of* *positive economic science*. Although these arguments are less of a challenge to the conventional wisdom, they still constitute a potentially significant change in the relationship between economics and moral philosophy.

Many of the arguments for increasing the (ethically) normative content of economic science come from the experimental literature, either experimental economics or experimental psychology. Researchers in these fields often reaches similar conclusions about the behaviour of the agents they study, although they differ regarding experimental protocols (particularly the role of cash payments) and how such results are to be interpreted (as a critique of rational choice theory or as a critique of the standard assumptions of rational choice theory). One of the systematic results of the literature has been that moral beliefs matter to decision making in experimental environments, and are sufficiently important that such morality often provides better empirical predictions than self-interested rational choice. For example, one of the earliest counter-intuitive experimental results was the tendency for individuals to over-contribute to (that is, not free ride on) public goods (Isaac et al. 1984). One explanation for this over-contribution is an ethical 'taste for fairness'. Another example involves the 'ultimatum game', a game where a self-interested rational agent should offer the smallest possible amount to the other player. The experimental evidence indicates that individuals do not generally behave as rational choice theory suggests, but rather give the other player a more 'fair' distribution. Since rational choice theory allows for the possibility of 'moral' (or otherwise non-self-interested) preferences, these results do not constitute a direct falsification of the core theory of rational choice (Guala 2005), but they certainly do challenge profession's traditional view of the positive and the normative. Instead of ethical norms interfering with the scientific investigation, these are cases where including ethical beliefs in the analysis improves the theory's descriptive accuracy.

The next two developments shift attention away from the positive–normative dichotomy but still challenge key features of the view passed down from Robbins and the ordinal revolution. According to the standard history of demand/choice theory, three (good) things happened as the theory of consumer choice progressed from the hedonistic cardinalism of the late 19th century, through the ordinal revolution of the 1930s, and on to the revealed preference/consistency interpretation in contemporary textbooks. First, all vestiges of hedonistic psychology were finally abandoned; second, all interpersonal comparisons of utility were eliminated; and finally, these changes brought about a steady improvement in the scientific foundations of the theory.

In recent years there has been serious reconsideration of at least two of these aspects of choice theory: hedonism and the impossibility of interpersonal utility comparisons. There have, of course, always been critics of the move away from hedonism and interpersonal utility comparisons (Harsanyi 1955; Robertson 1952), but the goal of such criticism has traditionally been to defend utilitarian ethics as the normative basis for economic policy. Appeals on such grounds certainly continue, but in recent years support for a return to hedonism and interpersonal utility comparisons has come from a number of new directions. Although these two topics are closely related, it is useful to discuss them separately.

Hedonism in rational choice theory is the idea that an agent's preference for a particular bundle of goods is based on the psychological feeling of satisfaction the agent receives when the bundle is purchased or consumed. This is clearly the notion of utility present in 19th century utilitarianism, and, even though it has been replaced by a non-hedonistic notion of preference in modern economics, it is still heard in casual conversation and in the classroom. One criticism of the move away from such psychological hedonism – a criticism from an earlier generation as well (Little 1957; Robertson 1952) – is that the move enervated the theory's ability to provide any real explanation of observed behaviour. Although this criticism has been a theme in a number of important recent studies (Davis 2003; Giocoli 2003; Mandler 1999), these authors do not generally recommend returning to a version of the earlier hedonist doctrine. On the other hand, some recent research does reach such neo-hedonist conclusions.

One research programme that endorses a return to hedonism is the work of the 2002 Nobel Prize winner in economics, the experimental psychologist Daniel Kahneman (Kahneman and Tversky 2000). Although the research of Kahneman and his associates is wide-ranging, and perhaps not every participant would support this particular aspect of the programme, the argument for a return to hedonism – what is called 'experienced utility' – has been a key aspect of Kahneman's approach (Kahneman 1994, 1999; Kahneman et al. 1997). There are two main parts to the argument for experienced utility, one philosophical and the other based on recent changes in our scientific tools. The philosophical argument is simply that weakening the positivist grip on experimental practice has opened the door to a number of new and fruitful possibilities; the more practical argument is that new tools for measuring experienced utility are becoming, and will continue to become, more available over time.

> The methodological strictures against a hedonistic notion of utility are a relic of an earlier period in which a behavioristic philosophy of science held sway. Subjective states are now legitimate topic of study, and hedonic experiences such as pleasure, pain, satisfaction or discomfort are considered open to useful forms of measurement. (Kahneman 1994, p. 20)

Paralleling such neo-hedonist arguments from experimental psychology are similar arguments from economics, particularly the literature endorsing 'happiness research' as a source of useful, and measurable, data for applied economic theory (Frey and Stutzer 2002). Economists appear to be more willing than psychologists to accept measures of happiness based on survey data, but the hedonistic themes are very much the same. Finally, there is a literature on the relationship between economic rationality and evolutionary biology that also suggests a hedonistic characterization of utility is scientifically appropriate (Robson 2001). It does not seem, as yet, that these newer interdisciplinary arguments defending hedonism have been integrated into the more traditional defence of utilitarian-based ethics as the basis for economic policy, but it is an obvious next step and is therefore extremely important for the relationship between economics and moral philosophy.

To turn from hedonism to a fourth change in the recent economics and ethics literature, there are similar (and often overlapping) arguments endorsing the revival of *interpersonal utility comparisons* in economics. Although the two issues – hedonism and interpersonal comparisons – are closely related, it is important to keep them separate. Hedonism is about feelings of pleasure and pain, and interpersonal comparisons are about having a common unit of comparison between the preferences of different agents (Mandler 1999). One can compare the current running through two different electrical appliances, but it is reasonable to conclude that such appliances do not 'feel' anything; similarly, two individuals could possess subjective, even cardinal, feelings about various goods and yet there would exist no way for a third party to measure or compare those feelings.

As in the case of hedonism, there have been consistent defenders of the legitimacy of interpersonal comparisons within economics, even when it was out of favour with most of the profession; many of these defenders came from the Marshallian tradition (Pigou 1920), but that is not exclusively the case (Harsanyi 1955, 1982). Often the argument was simply that economists should start with the observable facts of everyday life, and the fact is that humans make interpersonal comparisons all the time (Little 1957). Such defences continue, but in addition – again, as in the hedonism case – a number of new arguments are being made that draw on a range of interdisciplinary resources.

One source of evidence for interpersonal utility comparisons comes from recent research on neuroeconomics, part of the literature on 'the mind, the brain, rationality, agency and economics' discussed above. Neuroeconomics is a research programme that combines contemporary neuroscience and economics in the investigation of the microfoundations of decision making

(Glimcher 2003). Imaging studies from neuroeconomic research suggest that humans have the capacity to both represent the mental states of others and to empathize, that is, share the feelings of others. These abilities, it is argued, were selected for in human evolution because they 'enable people to predict others' behavior and, therefore, help them meet their individual goals' (Singer and Fehr 2005, p. 343). Neuroeconomics is not the only source of such arguments for the reliability, and survivability, of interpersonal utility comparisons. Similar arguments have also been made in the literature on the philosophy of mind. For example Alvin Goldman (1995) combines a reliabilist approach to the philosophy of science with various arguments from cognitive psychology to make the case for individuals having the ability to mirror, or simulate, the mental states of others in a reliable way, including interpersonal utility comparisons. In addition to the obvious support such research provides for moral theorizing within the utilitarian tradition, it also seems to provide a naturalistic explanation for the sympathy that played such an important role in Adam Smith's moral theory. At the very least, moral, economic and cognitive theorizing are simply different parts of a single intellectual exercise – as they were for Smith and Mill – rather than being hermetically isolated, as they were for most of the 20th century.

The fifth and final research to examine carries us outside the boundaries of the previous topics. Whether one is considering rational choice theory as normative theory, using moral preferences to explain observed behaviour in experimental economics, or defending hedonistic psychology and interpersonal utility comparisons, the discussion continues to be broadly within the research programme that identifies welfare with the satisfaction (or feelings received from the satisfaction) of individual preferences. In all of these cases, regardless of how much the recent literature conflicts with the mainstream view on such matters, the bottom line is still that individuals have preferences (hedonistic or not) and the individual 'good' is to have those preferences satisfied. But not all moral and political philosophy, even all that involves economics, follows this tradition.

John Rawls's *A Theory of Justice* (1971) is arguably one of the most important books on moral philosophy of the 20th century; it, and the philosophical discussion surrounding it, set the stage for many of the changes discussed above. Although Rawls's theory of justice falls squarely within the contractarian tradition – defining 'justice' as a property of the social contract that would emerge from the interaction of rational self-interested agents – he imposed strong restrictions on the context in which such contractual bargaining takes place; the decisions must be made in 'the original position' behind a 'veil of ignorance'. The principles of justice are those that would emerge from the bargaining of rational agents if those agents did not have any information about the position they would ultimately occupy (professional, class, gender, level of health, . . .) within the society governed by the contract, or even about what their preferences would be. Rawls goes on to argue for specific rules of justice that would emerge from such a context – including the much-debated 'difference principle' – but it is possible to separate his general approach to the question of justice from his specific distributional answers.

Although it is impossible to discuss the extensive literature surrounding Rawls's work in the space available here, it is important to consider the related contribution of one economist. The economist is Amartya Sen, the 1998 winner of the Nobel Prize in Economics. Sen has long been a critic of standard rational choice theory (Sen 1977), but his critical writings have come to be overshadowed by his own *capabilities* approach to social welfare and related issues (Sen 1985, 2002). The core idea of the capabilities approach to social welfare is to focus on the capabilities that people have, that is, on the things that people are effectively able to do or be – the functionings they are free to achieve – rather than on the satisfaction of individual preferences. Such capabilities are obviously multifaceted; they depend on the person's mental and physical characteristics as well as his or her social context and opportunities. One may have the capability to ride a bike, to find meaningful work, to express oneself artistically, or to participate in the governance of one's society; alternatively, one may have none, or only a few, of these capabilities. For Sen, such capabilities should be the proper focus for both the analysis of social welfare and the theory of economic development. The point of both welfare and development is to increase the capabilities of the population – to give them the freedom and opportunity to be better able to live the kind of life they find valuable. This, of course, does not rule out increasing the quantity of goods and services they have available, but it is at best only part of the story. In this sense Sen's approach actually moves us farther away from the traditional preference-based notion of social welfare than Rawls. Rawls's concept of justice is still based on the notion of a distribution of preference-satisfying goods (albeit primary social goods), while Sen shifts the focus away from individual preferences towards freedom and functioning.

Needless to say, Sen's approach has many critics, but his work has also generated an extensive supporting, extending and implementing literature. An important example of support and extension is Martha Nussbaum's (2000) research on women and development, which provides a specific list of the most important 'central human capabilities'; an example of implementation is the United Nations Development Program's Human Development Index, which builds on Sen's capabilities approach. Undoubtedly the capabilities literature will continue to evolve, but, regardless of the eventual shape it takes, it is an important contribution that has substantially changed the discourse on economics and moral philosophy.

Convergences

In closing, it is important to note the change that has taken place in the general way that various questions in philosophy and economics are approached in the recent literature compared with the way they were approached, at least by economists, for most of the 20th century. The traditional view considered 'the philosophical', whether it be epistemology or ethics, as something 'out there' with respect to economics. In the case of epistemology it was appropriate to seek methodological advice from philosophers about the character and practice of science, but the border crossing remained sporadic and one-way. In the case of ethics, the traditional view was simply to be aware of such ideas in order to prevent them from influencing the discipline's scientific practice.

Things have indeed changed. This is not to say that there is any consensus about specifics in the contemporary literature on either economics–epistemology or economics–ethics – in fact there has been an explosion of diversity and debate, and as such there is far less consensus on such matters than among economists in the past – but rather that the style of discussion has changed in both fields, and in a sense converged. Although a much longer list could be constructed, there seem to be three features of the debates in philosophy and economics discussed above that were effectively absent from the previous discussions: the interdisciplinarity, the naturalism, and the two–way relationship involved. The literatures discussed above all draw on a wide range of resources: economics and disciplinary philosophy certainly, but also cognitive psychology, neuroscience, the history and sociology of science, ideas from evolutionary biology, and a host of others. They are also broadly naturalist in focus in the sense that the relevant philosophical questions – whether epistemological or ethical – are on equal footing with the science, social or natural, that is employed in, and constrains, the philosophical discussion. Finally, and perhaps most obviously, work in philosophy and economics is much more of a two-way street. It is not simply that a shelf of scientific philosophy is 'applied' to economic methodology, or that a shelf of moral philosophy is used to cull normative concepts from economic science, but rather that economic notions of agency, choice, efficiency and equilibrium now condition the discussions in philosophy in the same way that alternative philosophical ideas, and 'normativity' more broadly, are increasingly involved in discussions within economic theory. On the one hand, these are substantive changes; on the other hand, such interconnections were present in the work of Smith, Mill and others. Perhaps these changes in the relationship between philosophy and economics are not so new after all; perhaps what needs explanation is not recent developments but the aberration of the 20th century.

See Also

- ▶ Conventionalism
- ▶ Epistemic Game Theory: Complete Information
- ▶ Ethics and Economics
- ▶ Experimental Economics
- ▶ Explanation
- ▶ Falsificationism
- ▶ Happiness, Economics of
- ▶ Instrumentalism and Operationalism
- ▶ Interpersonal Utility Comparisons (New Developments)
- ▶ Methodenstreit
- ▶ Methodology of Economics
- ▶ Positivism
- ▶ Scientific Realism and Ontology
- ▶ Theory Appraisal
- ▶ Value Judgements

Bibliography

Backhouse, R. 1997. *Explorations in economic methodology: From lakatos to empirical philosophy of science*. London: Routledge.

Blaug, M. 1980. *The methodology of economics: Or how economists explain*. 2nd ed. Cambridge: Cambridge University Press. 1992.

Blaug, M., and N. De Marchi, eds. 1991. *Appraising economic theories: Studies in the methodology of scientific research programs*. Aldershot: Edward Elgar.

Boland, L. 1997. *Critical economic methodology: A personal odyssey*. London: Routledge.

Caldwell, B. 1991. Clarifying Popper. *Journal of Economic Literature* 29: 1–33.

Caldwell, B. 1994. *Beyond positivism: Economic methodology in the twentieth century*. 2nd ed. London: Routledge.

Cartwright, N. 1989. *Nature's capacities and their measurement*. Oxford: Clarendon Press.

Cooter, R., and P. Rappoport. 1984. Were the cardinalists wrong about welfare economics? *Journal of Economic Literature* 22: 507–530.

Dasgupta, P., and P. David. 1994. Toward a new economics of science. *Research Policy* 23: 487–521.

Davidson, D. 2001. *Essays on actions and events*. 2nd ed. Oxford: Oxford University Press.

Davis, J. 2003. *The theory of the individual in economics*. London: Routledge.

Ferber, M., and J. Nelson. 2003. *Feminist economics today: Beyond economic man*. Chicago: University of Chicago Press.

Frey, B., and A. Stutzer. 2002. What can economists learn from happiness research? *Journal of Economic Literature* 40: 402–435.

Friedman, Milton. 1953. The methodology of positive economics. In *Essays in positive economics*. Chicago: University of Chicago Press.

Friedman, Michael. 1999. *Reconsidering logical positivism*. Cambridge: Cambridge University Press.

Giocoli, N. 2003. *Modeling rational agents: From interwar economics to early modern game theory*. Cheltenham: Edward Elgar.

Glimcher, P. 2003. *Decisions, uncertainty, and the brain: The science of neuroeconomics*. Cambridge, MA: MIT Press.

Goldman, A. 1995. Simulation and interpersonal utility. *Ethics* 105: 709–726.

Goldman, A., and M. Shaked. 1991. An economic model of scientific activity and truth acquisition. *Philosophical Studies* 63: 31–55.

Guala, F. 2005. *The methodology of experimental economics*. Cambridge: Cambridge University Press.

Hands, D. 1993. *Testing, rationality, and progress: Essays on the Popperian tradition in economic methodology*. Lanham: Rowman and Littlefield.

Hands, D. 1997. Caveat emptor: Economics and contemporary philosophy of science. *Philosophy of Science* 64: S107–S116.

Hands, D. 2001. *Reflection without rules: Economic methodology and contemporary science theory*. Cambridge: Cambridge University Press.

Harsanyi, J. 1955. Cardinal welfare, individualistic ethics and interpersonal comparisons of utility. *Journal of Political Economy* 63: 309–321.

Harsanyi, J. 1982. Morality and the theory of rational behaviour. In *Utilitarianism and beyond*, ed. A. Sen and B. Williams. Cambridge: Cambridge University Press.

Hausman, D. 1988. An appraisal of Popperian economic methodology. In *The Popperian legacy in economics*, ed. N. De Marchi. Cambridge: Cambridge University Press.

Hausman, D. 1992. *The inexact and separate science of economics*. Cambridge: Cambridge University Press.

Hausman, D., and M. McPherson. 2006. *Economic analysis, moral philosophy, and public policy*. Cambridge: Cambridge University Press.

Hoover, K. 2001. *Causality in macroeconomics*. Cambridge: Cambridge University Press.

Hutchison, T. 1938. *The significance and basic postulates of economic theory.* London: Macmillan.

Isaac, J., J. Walker, and S. Thomas. 1984. Divergent evidence on free-riding: An experimental examination of possible explanations. *Public Choice* 43: 113–149.

Kahneman, D. 1994. New challenges to the rationality assumption. *Journal of Institutional and Theoretical Economics* 150: 18–36.

Kahneman, D. 1999. Objective happiness. In *Well-being: The foundations of hedonic psychology*, ed. D. Kahneman, E. Diener, and N. Schwarz. New York: Russell Sage.

Kahneman, D., and A. Tversky, eds. 2000. *Choices, values, and frames*. Cambridge: Cambridge University Press.

Kahneman, D., P. Wakker, and R. Sarin. 1997. Back to Bentham? Explorations of experienced utility. *Quarterly Journal of Economics* 112: 374–405.

Kitcher, P. 1993. *The advancement of science: science without legend, objectivity without illusions*. Oxford: Oxford University Press.

Klappholz, K., and J. Agassi. 1959. Methodological prescriptions in economics. *Economica* 26: 60–74.

Lakatos, I. 1970. Falsification and the methodology of scientific research programmes. In *Criticism and the growth of knowledge*, ed. I. Lakatos and A. Musgrave. Cambridge: Cambridge University Press.

Latsis, S., ed. 1976. *Method and appraisal in economics*. Cambridge: Cambridge University Press.

Lawson, T. 2003. *Reorienting economics*. London: Routledge.

Little, I. 1957. *A critique of welfare economics*. 2nd ed. Oxford: Oxford University Press.

Maki, U., ed. 2001. *The economic world view: Studies in the ontology of economics*. Cambridge: Cambridge University Press.

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

McCloskey, D. 1998. *The rhetoric of economics*. 2nd ed. Madison: University of Wisconsin Press.

Mirowski, P. 2002. *Machine dreams: Economics becomes a cyborg science*. Cambridge: Cambridge University Press.

Mirowski, P. 2004. The scientific dimensions of social knowledge and their distant echoes in 20th-century American philosophy of science. *Studies in History and Philosophy of Science* 35: 283–326.

Mirowski, P., and E.-M. Sent, eds. 2002. *Science bought and sold*. Chicago: University of Chicago Press.

Mongin, P. 2006. Value judgments and value neutrality in economics. *Economica* 73: 257–286.

Morgan, M. 1999. Learning from models. In *Models as mediators*, ed. M. Morgan and M. Morrison. Cambridge: Cambridge University Press.

Morgan, M. 2001. Models, stories and the economic world. *Journal of Economic Methodology* 8: 361–384.

Nelson, J. 1996. *Feminism, objectivity and economics*. London: Routledge.

Nussbaum, M. 2000. *Women in human development: The capabilities approach*. Cambridge: Cambridge University Press.

Pigou, A. 1920. *The economics of welfare*. 1st ed. London: Macmillan.

Popper, K. 1965. *Conjectures and refutations*. 2nd ed. New York: Harper and Row.

Popper, K. 1968. *The logic of scientific discovery.* 2nd ed. New York: Basic Books.

Popper, K. 1994. Models, instruments, and truth: the status of the rationality principle in the social sciences. In *The myth of the framework: In defense of science and rationality.* London: Routledge.

Putnam, H. 2002. *The collapse of the fact/value dichotomy and other essays*. Cambridge, MA: Harvard University Press.

Rawls, J. 1971. *A theory of justice*. Cambridge, MA: Harvard University Press.

Robbins, L. 1952. *An essay on the nature & significance of economic science*. 2nd ed. London: Macmillan. (2nd edn first published in 1935).

Robertson, D. 1952. *Utility and all that and other essays*. London: Allen and Unwin.

Robson, A. 2001. Why would nature give individuals utility functions? *Journal of Political Economy* 109: 900–914.

Rosenberg, A. 1992. *Economics – Mathematical politics or science of diminishing returns?* Chicago: University of Chicago Press.

Ross, D. 2005. *Economic theory and cognitive science*. Cambridge, MA: MIT Press.

Ruccio, D., and J. Amariglio. 2003. *Postmodern moments in modern economics*. Princeton: Princeton University Press.

Samuelson, P. 1963. Problems of methodology – Discussion. *American Economic Review* 53: 231–236.

Searle, J. 2001. *Rationality in action*. Cambridge, MA: MIT Press.

Sen, A. 1977. Rational fools: A critique of the behavioral foundations of economic theory. *Philosophy and Public Affairs* 6: 317–344.

Sen, A. 1985. *Commodities and capabilities: The professor Dr. P Hennipman lectures in economics*. Vol. 7. Amsterdam: Elsevier Publishing.

Sen, A. 2002. *Rationality and freedom*. Cambridge, MA: Harvard University Press.

Sent, E.-M. 1998. *The evolving rationality of rational expectations*. Cambridge: Cambridge University Press.

Singer, T., and E. Fehr. 2005. The neuroeconomics of mind reading and empathy. *American Economic Review* 95: 340–345.

Weintraub, E. 2002. *How economics became a mathematical science*. Durham: Duke University Press.

Wible, J. 1998. *The economics of science: Methodology and epistemology as if economics really mattered.* London: Routledge.

Yonay, Y. 1998. *The struggle for the soul of economics*. Princeton: Princeton University Press.

Physiocracy

G. Vaggi

Keywords

Advances; Agrarian capitalism; Agriculture; *Avances annuelles*; *Avances foncières*; *Avances primitives*; Baudeau, N.; Cantillon, R.; Chalmers, T.; Circular flow; Circulating capital and fixed capital; Circulation of commodities; Civil society; Class; Colbert, J.-B.; Diderot, D.; Du Pont de Nemours, P. S.; Ephémérides du citoyen ou chronique de l'esprit national; Fixed capital; Forbonnais, F. V. D. de; Free trade; Galiani, F.; Gournay, Marquis de; Gray, J.; Hoarding; Input–output analysis; Laissez-faire; Le Trosne, G. F.; Luxury; Mercier de la Rivière, P.-P.; Mirabeau, V. R., Marquis de; Natural laws; Net product; Pattullo, H.; Peasants; Petty, W.; Physiocracy; Population growth; Productive and unproductive labour; Property rights; Quesnay, F.; Rent; Reproduction schemes; Single tax; Smith, A.; Spence, W.; Sraffa, P.; Surplus; *Tableau économique*; Turgot, A. R. J.; Underconsumptionism; Vauban, S. le P.; Voltaire, F. M. A. de; Wealth

JEL Classifications

B1

The Physiocrats lived and worked in France in the middle of the 18th century. The name derives from the title of a collection of some of the most important writings of their master François Quesnay, *Physiocratie, ou constitution naturelle du gouvernement le plus avantageux au genre humain* published in 1767 by P.S. Du Pont de Nemours. The term Physiocracy indicates the importance ascribed by these authors to natural forces, and derives from the Greek: *phýsis,* nature, and *kràtos,* power. The Physiocrats can be regarded as the first school of economists. They acted as an organized group of thinkers who intended to influence the French government's economic policy. They were accused of being sectarian because of their strict allegiance to the economic theories and opinions of their master, Quesnay. He provided the most important and original ideas, Victor Riqueti, Marquis de Mirabeau, was his first disciple, and included among the best known Physiocrats were Du Pont de Nemours, l'Abbé Nicolas Baudeau, Le Mercier de La Riviére and François Guillaume Le Trosne. One should also mention Henry Pattullo, an Irishman, who was deeply influenced by Quesnay's early articles (see Hecht 1958, vol. 1, p. 257). These French authors can be regarded as the 'inner circle' of the Physiocrats.

Another group of writers, sometimes confused with the Physiocrats, was Vincent de Gournay and his followers, the most famous of whom is Turgot. Gournay was appointed *Intendant* of commerce in 1751, and, like Quesnay, favoured laissez-faire. However, Gournay and his school never followed the Physiocratic programme and, in particular, disagreed on such important points as the idea that agriculture was the only productive sector of the economy.

Physiocracy covers a period of 20 years, from 1756 when Quesnay published his first economic articles in the *Encyclopédie* of Diderot and D'Alembert, until 1777 when Le Trosne's book appeared.

After a period of relative prosperity at the end of the 17th century and the beginning of the 18th century France experienced many bad years, mainly due to the backwardness of her agriculture. Often the Physiocrats recalled the age of Sully, Prime Minister to Henry IV, as the golden period of French agriculture and of the whole country. But now farmers were poor and could not implement the best methods of cultivation; the fiscal system was inefficient and unjust; and there were many different taxes and duties, both on the peasants themselves and on their products (Loménie 1879, vol. 2, p. 218). For instance, one had to pay an excise in order to take products from one province to another. Trade in agricultural products was greatly hindered by these impediments to the free circulation of commodities. There were also taxes that were levied on the number of people in the family – the various forms of *capitation*.

On top of these duties there were taxes which had to be paid to the Church and to the King, the *dîme* for the Church and the *taille* for the government. These taxes were levied on the revenue of lands, but their collection was extremely inefficient. The government used to sell the right to collect the *taille* in one province to some wealthy people who became tax collectors. This was the system of the *ferme général*, and was opposed by the Physiocrats because the peasants were oppressed by the *fermiers généraux*, who, having paid the government in advance, tried to make as much money as possible. They were allowed to keep all the taxes for themselves, and thus the King received much less money than that paid by the peasants.

There was a huge public deficit, and at the same time the peasants and the farmers were deprived of the fruits of agriculture.

The fiscal systems and the various barriers to the domestic and foreign trade of agricultural products discouraged the farmers from improving farming and agricultural productivity. During the first half of the 18th century there were many years of misery and famine (see Meek 1962, p. 46). According to Quesnay, during that period the population of France decreased from 24 million to 16 million (INED 1958, vol. 2, p. 506). He was too pessimistic (Eltis 1984, p. 39), but certainly French agriculture was unable to sustain a growing population. The Physiocrats compared the farming conditions in France with those in England, where farmers were rich and productivity was very high (INED 1958, vol. 2, pp. 440–41). The backward economic situation in France was made worse by the almost continuous wars, which absorbed human and financial resources. The Physiocratic movement must be examined in the light of this situation of recurring economic crises. The purpose of Physiocracy was to bring changes to certain characteristics in the French economy and in the political system of the *ancien régime*. They were a group of reformers, who tried to convince the rulers and the sovereign that some changes were needed to make the country more wealthy and politically stronger.

The history of the school can be divided into three periods: the years in which the main ideas appeared from 1756 to 1760, mostly in the works of Mirabeau and Quesnay; from 1760 there was a period of almost three years silence; the third period, from 1764 to 1777, saw a flourishing of writings and enterprises, thanks to the younger Physiocrats. Quesnay published his first economic articles in the *Encyclopédie,* to which he had been asked to collaborate on matters of agrarian economics. In 1756 *Evidence* and *Fermiers* appeared and in 1757 he published *Grains.* These works present most of the new ideas of the school and in particular they stress the view that agriculture is the most important sector in the economy. This is the corner stone of the Physiocratic theory of the nature and causes of national wealth. Quesnay identified the entire social product with the annual output of agriculture, and maintained that neither industry nor trade could increase the country's wealth, a doctrine which won him many enemies. In 1757 Quesnay met his first disciple: the Marquis de Mirabeau. This member of the French aristocracy had become famous because of his book *L'ami des hommes, ou traité de la population,* in which he stated that the wealth of a country depended upon the size of her population; like the title of his book he was called 'the friend of mankind', because of his liberal and reformist views. On July 1757 Quesnay and Mirabeau met at Versailles, where Quesnay was one of the King's physicians. Quesnay convinced Mirabeau that the products of land were more important than people because they secured the survival of the peasants and their families, who had to be regarded as the most important element in the economy (Weulersse 1910, vol. 1, pp. 55–6). Mirabeau was won over to the cause of Physiocracy, and in a couple of years he wrote many important works. In 1758 Quesnay wrote his famous *Tableau économique,* which was printed in three different editions between the end of 1758 and the first months of 1759. The analytical structure of Physiocratic economics was an enormous step forward. French society was divided into three main classes: the landlords – including the King and the Church – the farmers, and finally the artisans; the last two groups were respectively in charge of agricultural and industrial production. The *Tableau* outlines the main features of the process of circulation of commodities, at the end of the productive process, and gives a precise definition of the means of production and the net product. To illustrate his main economic ideas Quesnay used some rather obscure diagrams, which nevertheless greatly impressed the Versailles aristocrats. To make the *Tableau* more understandable to the public Mirabeau wrote some explanation in three further books of his *L'ami des hommes,* which were published between 1758 and 1760, and in which Quesnay's influence is very strong. Always in strict collaboration with the master, Mirabeau wrote a treatise on one of the major economic problems of the time: the reform of the fiscal system. The *Théorie de l'impôt* appeared in 1760 and presented one of the Physiocrats' most famous proposals: the single tax on rent. Fiscal reform must abolish all taxes and duties which are levied either on the peasants or on their products. This tax burden is one of the main reasons why cultivation cannot become profitable. The financial needs of the Kingdom must be met by a single general tax, which has to be paid in proportion to the net product of agriculture. This recommendation was the logical consequence of Quesnay's division of the social product into two parts: capital and surplus. The capital consists in the *avances* for farming, and must be preserved to maintain the same level of agricultural output. Any form of taxation falling on farmers' advances, *les avances*, would reduce the amount of capital employed in agriculture, and this would have disastrous effects on the whole country. Thus, only the surplus is really disposable for taxation, because it does not affect reproduction of output.

The largest part of agricultural net product accrued to the landlords in the form of rent. Quesnay and Mirabeau's single tax on the net revenue meant the abolition of all the fiscal privileges of the ruling classes, the Church, and the aristocracy.

Mirabeau and Quesnay tried to convince the nobles that in the following years their rents, net of taxes, would be much higher than before. In fact, the farmers, freed from the previous fiscal burden, would invest more money in the cultivation of land. The productivity of agriculture would rise, as would the surplus. But these arguments did not impress the nobility. Moreover, Mirabeau also violently attacked the tax collectors. The state must collect its taxes without the intermediation of these merchants and businessmen. But for many members of the aristocracy and the merchant bourgeoisie the role of tax collector meant power and wealth. Their reaction to Mirabeau's book was so strong that he was imprisoned for a few days, and then exiled to his countryside estate for some months (Loménie 1879, vol. 2, p. 226).

Here ended the formative period of Physiocratic school. Quesnay and Mirabeau did not publish anything for two and a half years. Du Pont wrote that Mirabeau's misfortune delayed the development of enlightenment (Du Pont, 1769, *Ephémérides*, vol. 2). Quesnay and Mirabeau spent this period of silence working towards a new book which was to be a fundamental text for Physiocratic doctrine. It appeared at the end of 1763 in three volumes, with the title *Philosophie rurale ou économie générale et politique de l'agriculture.* 1763 saw renewed interest in Physiocratic ideas; the government accepted the principle of free trade for corn inside France, which was one of the main reforms advocated by the Physiocrats. New followers joined Quesnay and Mirabeau; Du Pont de Nemours became an enthusiastic propagator of Physiocracy and in 1764 published a pamphlet in favour of free foreign trade for French corn. The mid-1760s were the period when Physiocracy had most influence on French economic policy.

In 1764 Du Pont became chief editor of a famous periodical, the *Journal de l'agriculture, du commerce et des finances*, which became an important vehicle for Physiocratic propaganda for some years. In the same year two new followers joined the school: Le Trosne and the less famous Saint Péravy. In 1765 Mercier de La Rivière was converted to Physiocracy. The school was now powerful enough to try to gain more influence on political and economic matters. After six years during which he mostly collaborated with Mirabeau's work Quesnay wrote again on his own, and from 1765 to 1768 he published many important articles intended to explain the principles of Physiocracy further, and to defend them from growing attack. At least three articles must be mentioned: ‘Le droit naturel’, which was written in 1765, the ‘Analyse de la formule arithmétique du Tableau économique’, which was written in 1766, both of which appeared in the *Journal de l’agriculture.* The latter is particularly important because it provides an easy explanation of the *Tableau économique* and in fact became its best-known version. The third work is the ‘Dialogue sur les travaux des artisans’, published in 1767 in the *Ephémérides.* Here Quesnay defended his view that only agriculture was capable of yielding a net product, while industrial activity was sterile because it only replaced the value of the raw materials and necessaries which had been used up in production.

During this period other Physiocrats contributed to the development of the school. In 1767 Mercier de La Rivière published his book *L’ordre naturel et essentiel des sociétés politiques*, in which he elaborated the political doctrines of Physiocracy. In the same year Du Pont published a collection of some of Quesnay’s work, entitled *Physiocratie,* where this term appears for the first time. The Physiocrats met every Tuesday in Mirabeau’s palace, and became a political group (Weulersse 1910, vol. 1, p. 132).

The abbé Baudeau too became a Physiocrat. In 1767 Baudeau founded an influential periodical, the *Ephémérides du citoyen* in which several Physiocrats collaborated. In the same year Du Pont started losing power in the *Journal de l’agriculture.* Since it was important for the Physiocrats to publish in a friendly periodical, they tried to win over Baudeau. After a few months of discussion the *Ephémérides* became the official periodical of Physiocracy. Many powerful people looked favourably on this group of intellectuals. Among them were Traudaine de Montigny and, above all, Turgot.

Physiocracy was also exerting some influence abroad. Catherine II invited Mercier de La Rivière to St. Petersburg to spread the new ideas. The Margrave of Baden also became a Physiocrat and exchanged letters with Du Pont. At home the Physiocrats had good relationships with the *encyclopédistes*; Diderot personally admired Mercier, but never shared the Physiocratic opinion that the wealth of a country derives from agriculture. The school also received support from the *Sociétés d’agriculture*, coalitions of wealthy farmers who tried to defend their interests and gain power over landlords. For these bourgeois farmers the doctrines of the Physiocrats were a powerful instrument of propaganda and political influence on the government.

The growing prestige and power of the Physiocrats also gained them new enemies including many of the aristocrats, and all those merchants who had exclusive trading privileges granted by the government.

In 1767 and 1768 many authors wrote against the Physiocrats. Grimm, Forbonnais and Mably, a disciple of Rousseau, attacked different aspects of Physiocracy. In his pamphlet *L’homme à 40 écus,* Voltaire ridiculed the Physiocrats’ fixation with numerical examples. The *encyclopédistes* became less friendly towards the Physiocrats. Some critics accused Quesnay and his disciplines of trying to mitigate the most unjust aspects of the *ancien régime* and to improve its inefficiencies only to prevent any major change in the French political system.

One particular point of Physiocratic theory came under attack at the end of the 1760s: the doctrine of the exclusive productivity of agriculture and the sterility of industry. (Notice that the Physiocrats regarded as productive not only the cultivation of soil but all the activities directly connected with agriculture, such as ‘grasslands, pastures, forests, mines, fishing’ (Kuczynski and Meek 1972, p. i).) Nobody questioned the importance of agriculture, but the attacks focused on the view that trade and above all industry were regarded as sterile occupations. This was the crucial point of the Physiocrats’ definition of wealth, and all their policy measures depended upon this doctrine. The liberalization of the corn trade, the reform of the fiscal system, and the attack on expenditures on luxury goods all depended upon the Physiocrats’ identification of national wealth with agricultural production. Many contemporary authors rejected the idea that national wealth could only be increased through land.

Veron de Forbonnais, a former pupil of Gournay, defended the productive role of commerce and industry (see Weulersse 1910, vol. 1, pp. 121–2). The strongest attack on the doctrine of the exclusive productivity of agriculture came from a Neapolitan priest, the abbé Ferdinando Galiani. With the help of Diderot, at the end of 1769, he published the *Dialogues sur le commerce des bléds,* in which he used brilliant prose to ridicule the supposed superiority of agriculture over industry. Galiani gave very simple and straightforward examples to show that increases in productivity were much more likely to take place in industrial production than in agriculture. Good and bad weather does not influence the output of manufacturing, and the advantages of the division of labour which derive from increases in capital stock are not limited by the existence of a fixed amount of soil (see Galiani 1770, p. 142). The decisive element which undermined the influence of Physiocratic views on government policy was growing opposition to the deregulation of the corn trade. From 1763 the commercial policy for the products of land, and in particular corn, had become one of the main economic issues in French society. We have already seen that the Physiocrats had some success with the 1763 declaration of the free circulation of corn inside France. In July 1764 an edict authorized the exportation of corn under certain circumstances. According to Quesnay laissez-faire in domestic and foreign trade was designed to favour the circulation of corn and increase its demand. A free trade policy implied the abolition of all the rights and the rules which hampered the corn market (Mirabeau 1764, vol. 2, p. 343). The merchants and all the people who had been granted some ‘exclusive privileges’ in corn trade were damaged, because they lost the position as middlemen between consumers and producers (INED 1958, vol. 2, p. 532). The Physiocrats wanted to favour direct contact between consumers and cultivators. The final outcome of a laissez-faire policy would have been an increase in the price received by farmers without damaging consumers, thanks to the squeeze, or the abolition, of the earnings of all intermediate agents. Free corn exportation would have further contributed to sustaining its demand and its price on the French market. The establishment of a *bon prix* for primary commodities was meant to raise farming’s profitability (INED 1958, vol. 2, p. 529). Farmers would have been able to make new investments, the productivity of French agriculture would have risen and the gross and net output of the primary sector would have been larger than before. This was the Physiocratic road to welfare and prosperity for the French Kingdom (Mirabeau 1760, vol. 2, p. 143).

In the second half of the 1760s the price of corn rose, but unfortunately this happened both in the wholesale and in the retail markets. It is difficult to ascribe this rise to free corn exports: most likely the price increases were due to a series of bad harvests. But the Physiocrats were accused of having contributed to the worsening of the living conditions of the French people, for whom corn was the most important consumption good. In 1768 there were popular uprisings against the high price of corn both in Paris and in the countryside.

Part of public opinion began to consider Physiocratic theory as a dangerous attack on poor people, and some parliaments, in particular those of Paris and Rouen, called for the reintroduction of the restrictions on the corn trade. Between 1768 and 1770 there were many discussions for and against laissez-faire for primary commodities; the public and the government itself came gradually to oppose the Physiocratic views. At the end of 1769 the abbé Terray, one of the fiercest opponents of Physiocracy, was appointed *contrôleur général*, a sort of minister in charge of all economic matters. There were more uprisings and more declarations by provincial parliaments against the free exportation of corn.

During this period relationships between the Physiocrats and the *encyclopédistes* deteriorated notably. Grimm made fun of the *secte* of the *philosophes économistes,* who were accused of presenting a reactionary doctrine, designed to favour the landlords and the rural classes against the people in the cities (Weulersse 1910, vol. 1, pp. 230–1). In this hostile climate Turgot and Morellet rejected the invitation to join the Physiocrats. After a period of irregular publication the *Ephémérides* were put under censorship. At the end of 1770 corn trade legislation was completely changed and strict regulations were introduced both in foreign and domestic trade. The period of political influence of Physiocracy was almost over, and Physiocratic doctrines rapidly disappeared from public debate and the political arena. A final glimpse of the Physiocrats' impact on French economic policy was due to Turgot. On becoming *contrôleur général* in 1774 he restored internal free trade in corn, with the exception of Paris (Groenewegen 1977, p. xxxii). But this policy had many powerful opponents, and caused Turgot's fall after two years.

At the end of the 1760s Physiocracy was on the wane. After 1768 Quesnay wrote nothing on economic matters, lost interest in economic problems, and spent the last years of his life studying geometry; he died in 1774. In the 1770s there were only two works by Physiocrats. In 1772 Du Pont published an *Abrégé des principes de l'économie politique.* In 1777 Le Trosne published his book *De l'intérêt social, par rapport à la valeur, à la circulation, à l'industrie et au commerce intérieur et extérieur*. These attempts to renew interest in Physiocracy failed to influence either the policies of the government or discussions in French society.

Quesnay and his followers must be regarded as part of that cultural phenomenon which was the French Enlightenment. Many authors had already pointed out France's disastrous economic circumstances in the first half of the 18th century. Moreover, most of these writers did not limit their denunciations to the unjust and inefficient aspects of French society, but extended their investigations to the problem of the origin and nature of civil societies and the analysis of the best rules and laws which should regulate the relationships between individuals.

Quesnay and his disciples contributed to the French enlightenment. They concentrated their efforts on the economic and social reforms which were needed to make France more efficient. But they were not very much interested in the analysis of the fundamental principles of the civil societies and the role of subjects and the state; such issues have no prominence in their works. The only exception is the book by Mercier de La Rivière, which is mainly dedicated to the analysis of the political system. In general the Physiocrats never questioned the existence of the absolute monarchy and the political organization of the *ancien régime.* This is one of the main reasons why they were accused of being too hesitant in the defence of individual rights against state power. Montesquieu's *L'Esprit des lois* was one of the philosophical works which most influenced the Physiocrats; this book appeared in 1748. A year later Rousseau published his *Discours,* with which the Physiocrats were much less in agreement.

In the first half of the century, several authors had already analysed the economic and social conditions in France, paying particular attention to the agricultural sector. In different ways these writers can be considered as the forerunners of Physiocracy. We have already seen that at the beginning of the 17th century, French agriculture was prosperous thanks to Sully, Henry IV's Prime Minister. The years of Louis XIV were marked by Colbert's attempt to favour industrial activities by keeping the prices of subsistence goods low. At the end of the 17th century there was a reaction to Colbert's policy and the role of agricultural production was again emphasized. Among the authors who influenced the Physiocrats Vauban and Boisguillebert should be recalled. In 1707 Vauban published *Dîme royale*; he said that a single tax on agricultural output was the best solution to France's fiscal problems.

In 1695 Boisguillebert published the *Détail de la France,* a collection of statistical information on the French economy, and in 1707 his *Dissertations sur la nature des richesses, de l'argent et des tributs* appeared. He stressed the importance of agriculture among various economic activities, and above all described the production and exchange of commodities in terms of a circular flow, a sort of selfregenerating circuit. Two other works which deserve mention are Melon's *Essai politique sur le commerce* (1734) and Herbert's *Essai sur la police des grains* (1754). But of the writers who exerted a major influence on Physiocracy, a special place is occupied by Cantillon. Other British authors were well known in France in those days, for instance

Child, Tucker and Hume, but Cantillon's impact on Physiocratic theory was much deeper.

At the beginning of the 1740s Mirabeau had a copy of Cantillon's *Essai sur la nature du commerce en général,* which he regarded as the fundamental text on economic matters, an opinion shared by Quesnay. Mirabeau published Cantillon's *Essai* in 1755. In many ways the Physiocrats followed Cantillon's general approach to economic analysis. Cantillon gave a framework in which to build up a theoretical model of the working of the whole economic system. Economics would no longer be a subject for pamphleteers, merchants and practical men, but would become a topic of separate theoretical speculation. Practical matters would be examined in terms of the theories' fundamental principles. Cantillon's analysis left clear marks on Physiocratic economics, such as, for example, his classification of the people of a kingdom into three main classes: landlords, entrepreneurs and workers. His analysis of the distribution of income was related to these classes; he spoke of the farmers' 'three rents', which make up the value of the products and each rent is the income of one class (Cantillon 1755, p. 43). Like the Physiocrats Cantillon emphasized the productive role of the farmers as entrepreneurs. Finally, Cantillon considered expenditures of revenue as the most important element in the determination of the prices of commodities and the level of activity of the sectors other than agriculture. It is thanks to the landlord's expenditures of their revenues that these activities can exist. Through Cantillon the Physiocrats were also influenced by Sir William Petty.

With regard to Physiocratic theory it must be remarked that almost all the main contributions are due to Quesnay. As to their philosophical views they believed that civil societies were only a mirror of natural order. The Physiocrats believed that societies are characterized by the existence of laws which govern the relationships between individuals. These natural laws can be studied, and their knowledge provides the foundation for the proper administration of the country. It must be noticed that the Physiocrats' attitude towards natural laws and natural order is somewhat different from most of their French contemporaries and from Adam Smith. Natural laws operate quite independently of men's will (INED 1958, vol. 2, p. 526), but at the same time they are not so powerful as to be ignored. These laws have been inscribed in nature by God himself (Mirabeau 1764, vol. 2, pp. 9–11; INED 1958, vol. 2, p. 934), but their working can be hampered and their effects can be modified by unwise ruling of society and by powerful social groups. Therefore natural laws do not necessarily overwhelm men's actions, and civil societies cannot be analysed as if they were a mechanical system which always gives the same results. The Physiocratic concept of natural order is a peculiar mixture of objective laws and of socio-historical modifications. Natural laws exist, and can be studied and precisely singled out, but there is also room for active human intervention. This view of the natural order has far reaching implications. The Physiocrats believed that societies evolve through definite specific stages (Meek 1976, pp. 72, 99). But this evolutionary process can be stopped for long periods. They regarded England as the country where natural law displayed its positive effects, and which reached the highest stage of economic development. However, in France civil laws and historical traditions prevented the full unfolding of natural laws and the country was still in a backward condition. Thus, the Physiocrats did not take a deterministic approach to the study of societies, even if they believed in the existence of objective natural laws. Natural order is a sort of normative situation which describes the features of an ideal society.

How can these natural laws be discovered? Quesnay wrote an article entitled *Evidence* in which he maintained that the laws of natural order reveal themselves in day-to-day events. The Physiocrats were also influenced by Descartes, a fact which helps to explain their belief in knowledge through evidence. In some way natural laws seem to be inborn in men, and this is why the system of natural order should be clear to everyone.

Which are the fundamental principles of natural order? Here too the Physiocrats' answer shares some features of contemporary French culture but also presents some peculiarities. Quesnay and

Mercier de La Rivière contributed to the development of the philosophical and political views of the school. In 1765 Quesnay wrote *Le droit naturel,* where he mentions the natural rights of men, which, however, are discussed mainly in relation to the economic features of society. Thus, freedom implies the abolition of privileges and regulations in all markets. Free competition must rule in the labour market as well as in domestic and foreign trade; people must be entirely free to decide how to spend their revenues. For the Physiocrats freedom meant universal competition, and was regarded as the basis for the increase in private and public wealth.

They regarded private property as a fundamental right of men, but by this they meant that land ownership was part of the natural order of societies, and the King was considered to be co-owner of all French soil. But the Physiocrats also emphasized the importance of guaranteeing the farmers and their families the ownership of the capital employed in agriculture and the fruits of farming.

Private ownership excludes the possibility of equality between men; indeed development of the economy will cause more inequalities. Differences among people are necessary in order to have an efficient economic system, capable of yielding a high net product. Therefore the political structure of a country reflects its economic and social circumstances. For the Physiocrats the major forces which explain historical changes in societies must be sought in their economic structure. This economic interpretation of history (Meek 1962, p. 376) underlines the fact that economic systems are based on the existence of different social groups which have separate economic functions. The Physiocrats distinguish French subjects into three classes: the landlords, including the King and the Church who represent the First and Second Estate; the people working in agriculture; and the industrial workers. This tripartite distinction is a hybrid since it is based partly on intersectoral differences and partly on property relationships. But in Physiocracy there is also a more detailed class analysis. In agriculture there are both farmer entrepreneurs and salaried peasants; with some ambiguity the same distinction between employers and employees exists in the industrial sector too. Then there are the merchants and all the people related to trade, and here too a whole class is identified with a sector of the economy. However unsatisfactory this approach may be, it was to be extremely important in the development of economic theory. First, following Petty, Boisguillebert and Cantillon the Physiocrats consider the economy as a system which is made up of different social groups, and which tends to reproduce both its economic and social relationships. Second, these classes are defined according to their role in the process of production and circulation of commodities. These two features are typical of the whole of classical political economy. Of course, the main limitation of the Physiocratic analysis of classes is the fact that they tend to identify social groups with the sectors of the economy, even if there are also hints of a distinction based on political and economic power relationships.

The Physiocrats' concept of natural order deeply affects their political views. In general they argue that the principles of political order must accomplish those of the natural order. The particular way in which this connection between the two orders comes about is through the form of government which they call *despotisme légal.* The supreme authority is that of the absolute hereditary monarchy which does not need to be legitimatized by the subjects. Of course this view was criticized by many authors of the time. According to the Physiocrats the only authority was that of the sovereign and that came directly from God. The King was also the natural owner of all the territory; *despotisme* was also patrimonial; the King was also the highest tutor of all forms of property. He was a legal despot, because he guaranteed security and freedom in property. Property is the key notion in the foundations of a political order. But according to the Physiocrats the authority of the King was moderated by the fact that he had to exert his power as an enlightened sovereign. By this the Physiocrats meant that the King had to be aware of natural laws and had to favour their implementation in civil society. The King's knowledge of natural laws was the decisive element which had to secure the existence of appropriate civil laws and just administration. No confrontation could exist between the sovereign and his subjects because the King, properly instructed about natural order, knew that his interests coincided with those of citizens.

The Physiocrats envisaged only two limits to the King's power. On the one hand there was public opinion, which was also instructed in the principles of natural order, so that the people could react to a situation where the King ignored natural laws. On the other hand the fact that the King was the owner of the whole country did not entail exploitation of his subjects. Customs and habits about the fiscal system could not be modified by the sovereign's own decision alone.

A final peculiarity of the Physiocrats' political thinking is their view that only agriculture produces a surplus. In an agricultural country like France the merchants and all the owners of monetary and financial wealth are not part of the nation because their interests are in opposition to those of the state. The only true citizens are the landowners, the wealthy cultivators, and the other people directly linked to agricultural production; the artisans of the industrial sector were somehow tolerated. It is clear that Physiocrats aimed at a political system based on the alliance of all social groups linked to agriculture and the King.

The distinctive feature of Physiocratic economics was the doctrine of the exclusive productivity of agriculture; only activities directly linked to nature could yield a net product over costs. To justify their views Quesnay and the Physiocrats used many different arguments. Agriculture was superior to other economic sectors because it produced the raw materials and the necessities for all other occupations. The subsistence of all people could only come from farming (INED 1958, vol. 2, p. 775). Industrial and commercial activities could exist only because the peasants were producing more foodstuffs than was required for their own subsistence. Moreover, France had been endowed by nature with a large and fertile territory and was surrounded by countries whose soil was much less suitable for farming and who were potential buyers of French products (Le Trosne 1777, p. 988; INED 1958, vol. 2, pp. 600–1).

The fiercest attacks by Physiocracy concerned the view that industrial and commercial activities were sterile. The Physiocrats believed that in all trading activities there was only an exchange of commodities of equal value, but these values had already been produced elsewhere.

All merchants and middlemen who operated in ‘resale trade’ were a burden to society, since they had to be maintained without adding anything to national wealth (INED 1958, vol. 2, p. 947; Mercier 1767, p. 278). The Physiocrats saw that some traders were making large monetary fortunes, but this was not a proof of their productiveness; on the contrary, this was the result of a violation of natural laws. Merchants could become rich thanks to unequal exchanges due to exclusive trading privileges. These regulations contradicted the natural principle of free and unobstructed competition in all markets. Industrial activities simply transformed the products of agriculture into different types of commodities, whose exchange values had already been determined (INED 1958, vol. 2, pp. 496, 865). The sterility of industry was then explained by the fact that, according to the Physiocrats, the value of its product was equal to the value of its expenses and there was no net product left.

The profitability of farming is the most important requirement for the accumulation of capital in agriculture. Hence commercial policy must be designed to sustain the exchange value of the products of land. Free trade was the main way to raise the prices of primary commodities and induce the farmers to reinvest their profits in farming (INED 1958, vol. 2, p. 602).

It is important to notice the Physiocrats considered laissez-faire instrumental in the establishment of favourable trading conditions for French farmers. Quesnay and his disciples were not in favour of a generalized free trade, and they were not particularly interested in the commercial conditions of manufactures; their only aim was the achievement of high exports of primary products. They looked to a positive balance of trade for French agriculture, since France should have become the granary of Europe. Moreover, Physiocrats regarded foreign trade as necessary only because the French domestic market was too small and too poor to guarantee the profitable sale of French corn (INED 1958, pp. 848–9). With a larger domestic market there would be no need to export corn.

Quesnay was quite aware of the important role of markets; a large consumption was necessary to sustain the prices of agricultural commodities. The Physiocrats believed that there was no lack of potential demand for corn, since it was a fundamental item. The main problem of French agriculture was not the lack of potential consumers but the lack of effective consumption (INED 1958, pp. 528, 963). The exchange value of corn was affected by the number of effective consumers and by their wealth (p. 824): these were the true causes which determine the price of corn. The demand of those people who were not rich enough to pay for corn at its proper price was of no interest for the economy, according to the Physiocrats. They argued that landlords should spend most of their revenues in the purchase of agricultural products to increase the effective demand for French foodstuffs. Landlords were the social class which received most of the surplus, as rent, and all activities depended on the expenditure of this revenue.

The Physiocrats noted that the way in which revenue is spent influences society's economic structure. For instance, if the landlords buy many primary commodities and few manufactures, agriculture grows at a faster pace than industry (Kuczynski and Meek 1972, p. 12). Of course the Physiocrats were in favour of high consumption of agricultural products which they called *luxe de subsistence,* and were against the purchase of industrial goods, *luxe de décoration* (Baudeau 1767, pp. 190, 217). The Physiocrats attacked luxury because they wanted to encourage the profitable sale of agricultural products. They also opposed savings and the hoarding of money which would end up in monetary stocks to be lent at interest (Mirabeau 1764, vol. 2, p. 343). Monetary and financial fortunes were not a true form of wealth, but represented a deduction from the process of circulation of agricultural commodities.

The concept of the net product is the Physiocrats' main contribution to economic theory. This notion is related to that of advances, a term they used to indicate the means of production. The social product must include all the goods which make up the advances, and for each of them the quantity produced must be at least equal to the quantity which has been used as input.

Physiocratic analysis of the different types of advances is the first classification of the means of production, or capital, in the history of economic theory. The *avances foncières,* or land advances, included all the operations necessary to prepare a piece of land for farming. *Avances annuelles* are another important type of advances, this time annual ones. They are made by farmers and consist of products which must be invested in cultivation at each productive cycle because they are completely consumed during the process of production. These commodities include raw materials and necessaries which allow the peasants and their families to work during the year, but some interpreters of Physiocracy maintain that they also include some manufactured goods (Meek 1962, pp. 274–5; Eltis 1984, pp. 29–31). Annual advances are a typical kind of circulating capital.

The original advances, *avances primitives,* are made up of instruments and equipment which last for more than one year; they also include livestock (Eltis 1975, p. 189). All these commodities must be regarded as fixed capital lasting for many years (INED 1958, vol. 2, p. 798). In fact Quesnay indicated that the average life cycle of the *avances primitives* lasted ten years. Productivity increases are closely related to capital accumulation (ibid., pp. 427 ff.). A prosperous economy is characterized by large-scale farming, where agriculture employs a large stock of *avances primitives.* This view of the ideal economic system has been called 'agrarian capitalism', since agriculture is the most advanced capitalist sector (Hoselitz 1968).

According to Quesnay, in ideal agricultural production the value of the fixed capital must be five times that of the annual capital and is assumed to be ten 'milliards'. Given an annual rate of decay of ten per cent, the farmer must repay a fixed amount of capital equal to a half of the circulating capital. Therefore, the overall *réprises,* or returns, which make up the value of all the means of production annually consumed is given by the sum of the whole annual advances plus '1 *millard livres*' depreciation of *avances primitives* (Meek 1962, p. 154).

At the end of the 1770s the political reactions to the economic policy suggested by the Physiocrats caused a decline in their intellectual influence. Echoes of Physiocratic economics survived in some European countries such as Russia, Poland, Germany, and Tuscany, and in the United States.

But Quesnay and his followers left important marks in the history of economic theory. At the end of the 18th century and at the beginning of the 19th century several British economists looked quite favourably on many ideas of the Physiocrats. In different ways, John Gray, William Spence and Thomas Chalmers defended the superiority of agriculture over industrial activities (Meek 1962, pp. 345 ff.). Because of the way in which they stressed the importance of demand and consumption in sustaining economic activities, the Physiocrats were also regarded as forerunners of underconsumption theories. Lack of consumption and the excess of expenditure on luxury goods could cause economic crises. Thus, the Physiocrats recognized the possibility of economic breakdown. From this point of view Physiocracy can be related to Sismondi and Malthus (Meek 1962, pp. 313 ff.). The *Tableau économique* does not only describe the necessary economic relationships between some economic magnitudes, but it can also indicate why and how the ideal conditions of production could break down. Quesnay himself provided several examples of *Tableau* 'in disequilibrium' (Eltis 1975).

The major merit of the Physiocrats is that of having given a fundamental contribution to the rise of that stream of thought which was classical political economy. They precisely defined the concepts of surplus and capital; they introduced the distinction between productive and sterile activities. The Physiocrats clearly distinguished the social classes according to their role in production. Therefore the Physiocrats can properly be regarded as the first inspiration of that economic theory which goes by the name of the surplus approach. In the *Theories of Surplus Value* (Marx 1864–5, vol. 2, ch. 2), Marx indicated the Physiocrats as the first authors who adopted this approach for the analysis of economic systems. One aspect of Marxian economics which is derived from Physiocracy is the description of the economy by means of reproduction schemes. It must be noticed that the first two sectors in Marx's reproduction schemes coincide with those of Quesnay, that is, agriculture and industry (Marx 1867–74, vol. 2, part 2).

The Physiocratic distinction between productive and unproductive labour can be found in all major classical economists, from Smith to Malthus and Ricardo, even though they gave different solutions to this problem.

The surplus approach, which was characteristic of classical economists and of Marx, was again brought to the fore in the 1960s thanks to Piero Sraffa's book *Production of Commodities by Means of Commodities.* Sraffa refers to the Physiocrats as one of his sources (see Sraffa 1960, appendix D). Physiocratic economics also influenced other aspects of modern economic theory. Leontief's input–output analysis finds an important forerunner in the *Tableau économique*, while the distinction between productive and unproductive labour has been the focus of renewed interest and has been used to investigate the failures of some modern economic systems (Bacon and Eltis 1976, preface).

The influence of the Physiocrats on Adam Smith deserves special attention. Smith was in France for three years between 1763 and 1766, and was in touch with some Physiocrats and with Turgot. Certainly Smith was well aware of the debates which were taking place during those years about Physiocratic economics, and it is generally admitted that he borrowed some specific concepts from Quesnay. These are the concepts of net product, its difference with the capital advanced, and the distinction between production and unproductive labour. These concepts did not appear in Smith's economic writings before his visit to France, but played an important role in the *Wealth of Nations.* The Physiocrats' influence on Smith is further proof of their important place in the building of classical political economy. In the *Wealth of Nations* Smith dedicates many pages to explain Physiocratic economics (Smith 1776, book 2, chapter 9). He criticized many aspects of Physiocracy; for instance, while accepting the idea that agriculture was the most important economic sector of the country, he did not agree that industry was sterile. For Smith many features of Physiocratic economics were not appropriate to explain the workings of modern commercial societies like England. Physiocracy was too influenced by the economic conditions of 18th-century France, and was thus particularly useful to study agricultural societies. But for Smith, Physiocracy was greatly superior to mercantilism and it was the necessary basis on which to found the new economic science, or as Smith wrote ‘the nearest approximation to truth’ (Smith 1776, vol. 2, p. 199).

See Also

- ▶ Du Pont de Nemours, Pierre Samuel (1739–1817)
- ▶ Ephémérides du citoyen ou chronique de l’esprit National
- ▶ Quesnay, François (1694–1774)

Bibliography

Bacon, R., and W. Eltis. 1976. *Britain’s economic problem: Too few producers*. London: Macmillan.

Baudeau, N. 1767. *Explication du Tableau économique à Madame de ***, par l’auteur des Ephémérides*. In *Ephémérides du citoyen*, vols. 11 and 12.

Beer, M. 1939. *An inquiry into physiocracy*. London: Allen & Unwin.

Boisguillebert, P.P.S. 1707. Dissertation sur la nature des richesses, de l’argent et des tributs. In *Economistes et financiers du dix-huitième siècle*, ed. E. Daire. Paris: Guillaumin, 1843.

Cantillon, R. 1755. *Essai sur la nature du commerce en général* (Ed. H. Higgs). London: Frank Cass, 1959.

Daire, E. (ed.). 1846. *Physiocrates*. Paris: Guillaumin.

Du Pont de Nemours, P.S. 1767. *Physiocratie, ou Constitution naturelle du gouvernement le plus avantageux au genre humain.* Leyden/Paris.

Du Pont de Nemours, P.S. 1772. *Abrégé des principes de l’économie politique.* In Daire (1846).

Eltis, W.A. 1975. François Quesnay: A reinterpretation. 1. The Tableau économique. *Oxford Economic Papers* 27(2): 167–200.

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

Galiani, F. 1770. *Dialogues sur le commerce des bléds*. Milan/Naples: Riccardo Ricciardi, 1959.

Groenewegen, P.D. 1977. *The economics of A.R.J. Turgot*. The Hague: Nijhoff.

Hecht, J. 1958. La vie de François Quesnay. In INED, vol. 1.

Hoselitz, B.F. 1968. Agrarian capitalism, the natural order of things: François Quesnay. *Kyklos* 21: 637–662.

INED (Institut Nationale d’Etudes Démographiques). 1958. *François Quesnay et la Physiocratie,* 2 vols. Paris: INED.

Kuczynski, M., and R.L. Meek. 1972. *Quesnay’s Tableau Economique*. London: Macmillan.

Le Trosne, G.F. 1777. *De l’intérêt social, par rapport à la valeur, à la circulation, à l’industrie et au commerce intérieur et extérieur.* In Daire (1846).

Loménie, L. 1879. *Les Mirabeau – Nouvelles études sur la société française au XVIII siècle*, vol. 2. Paris: Dentu.

Marx, K. 1864–5. *Theories of surplus value.* London: Lawrence & Wishart, 1963.

Marx, K. 1867–94. *Capital.* London: Lawrence & Wishart, 1970.

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

Meek, R.L. 1968. Ideas, events and environment – The case of the French Physiocrats. In *Events, ideology and economic theory*, ed. R.V. Eagly. Detroit: Wayne State University Press.

Meek, R.L. 1976. *Social science and the ignoble savage*. Cambridge: Cambridge University Press.

Mercier de la Rivière, P. 1767. L’ordre naturel et essentiel des sociétés politiques. In *Collection des économistes et des réformateurs sociaux de la France*. Paris: Geuthner, 1910.

Mirabeau, V.R. 1758–60. *L’ami des hommes, ou Traité de la population.* Aalen: Scientia Verlag, 1970.

Mirabeau, V.R. 1760. *Theorie de l’impôt*. Aalen: Scientia Verlag, 1972.

Mirabeau, V.R. 1764. *Philosophie rurale, ou Economie générale et politique de l’agriculture*. Aalen: Scientia Verlag, 1972.

Pattullo, H. 1758. *Essai sur l’amélioration des terres*. Paris: Durand.

Petty, W. 1662. A treatise of taxes and contributions. In *The economic writings of Sir William Petty*, vol. 1, ed. C.H. Hull. Cambridge: Cambridge University Press, 1899.

Smith, A. 1776. *An inquiry into the nature and causes of the wealth of nations* (Ed. E. Cannan). London: Methuen, 1961.

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

Voltaire, F.-M. 1767. *L’homme aux quarante écus.* In *Oeuvres complètes de Voltaire,* vol. 51. Paris: Baudouin, 1825.

Weulersse, G. 1910. *Le mouvement Physiocratique en France (de 1756 à 1770)*. Paris: Félix Alcan.

Pierson, Nicolaas Gerard (1839–1909)

Arnold Heertje

Keywords
Austrian economics; Cohen Stuart, A. J.; Edgeworth, F. Y.; Education; Pierson, N. G.; Taxation, pure theory of

JEL Classifications
B31

Born in Amsterdam, 7 February 1839; died in Heemstede, 24 December 1909. A Dutch economist of international reputation, Pierson dominated economics in the Netherlands during the second half of the 19th century. He started his career in the commercial and banking world of Amsterdam. He became President of the Dutch Central Bank, Minister of Finance and Prime Minister. As an economist he was a self-educated man, just like David Ricardo, but he was nevertheless invited to become Professor of Economics at the University of Amsterdam. He taught in the Faculty of Law from 1877 onwards until 1885. Broadly speaking, he advocated the main ideas of the Austrian school of thought in economic theory, although he maintained a material concept of welfare and production. On money, banking and taxation he was a well-known authority, who stimulated Cohen Stuart to write his famous dissertation on the application of utility theory to taxation. His knowledge of the history of ideas was outstanding and he was one of the first to recognize the significance of the Italian authors of the 17th and 18th centuries.

As a political economist, Pierson was basically in favour of a market economy. He was a critic of Marxism, but still not against a modest degree of state intervention. He advocated, in particular, the importance of high-level education, organized by the government, in order to improve the condition of the working class.

In 1863 he wrote a booklet on the future of the Dutch Central Bank, in which he strongly defended the monopolistic position of the Bank with regard to the creation of banknotes (Pierson 1863). An English translation of his very popular textbook also appeared (1902a).

Pierson's analysis of value in a socialist society (1902b) is of lasting significance.

See Also

► Cohen Stuart, Arnold Jacob (1855–1921)

Selected Works

1863. *De toekomst der Nederlandsche Bank*. Haarlem: Kruseman.

1902a. *Principles of economics*. London: Macmillan.

1902b. The problem of value in the socialist society. In *Collectivist economic planning*, ed. F.A. von Hayek. London: Routledge, 1935.

Bibliography

Edgeworth, F.Y. 1925. *Papers relating to political economy*. London: Macmillan, vol. 1, 351–356; vol. 3, 77–85.

Heertje, A. 1989. Nicolaas Gerard Pierson. In *Perspectives on the history of economic thought*, vol. 1, ed. D.A. Walker. Aldershot: Edward Elgar.

Hennipman, P. 1964. *Handwörterbuch der Sozialwissenschaften*. Tübingen: Mohr.

Van Maarseveen, J.G.S.J. 1981. *Nicolaas Gerard Pierson*. Rotterdam: Erasmus University.

Pigou, Arthur Cecil (1877–1959)

Nahid Aslanbeigui

Abstract
Arthur Cecil Pigou founded welfare economics by synthesizing Marshall's theoretical framework and Sidgwick's categories of market

failure and imperfections. His view of welfare economics was expansive, including resource allocation, income redistribution, business cycles, and unemployment. Pigou made important contributions to other areas of economics as well: the theory of value, public finance, index numbers, and evaluation of real national income. The most neglected aspect of Pigou's work is his investigation of a remarkable range of labour-market phenomena explored by subsequent economists – implicit contracts, internal labour markets, wage rigidity, labour market segmentation, human capital theory, and collective bargaining.

Keywords

Bandwagon effect; Business cycles; Cambridge school; Coase, R. H.; Collective bargaining; Cost–benefit analysis; Double taxation; Externalities; Factor prices; Fixed factors; Happiness; Harris–Todaro model; Ideal output; Increasing returns; Inheritance tax; Interdependent utility; Interpersonal utility comparisons; Involuntary unemployment; Kahn, R.; Keynesian revolution; Marshall, A.; Mathematical economics; Money; Money illusion; Monopoly; Natural monopoly; Natural rate of unemployment; Pigou, A. C.; Pigou effect; Positive economics; Price discrimination; Principal and agent; Public goods; Public works; Real balances; Real wages; Redistribution of income; Relative income; Robbins, L. C.; Robertson, D.; Robinson, A.; Robinson, J. V.; Sidgwick, H.; Snob effect; Social cost; Sraffa, P.; Stationary state; Sticky wages; Technical change; Trade unions; Utilitarianism; Welfare economics; Young, A. A.

JEL Classifications

B31; A1; B3; D6; J0

Arthur Cecil Pigou was founder of welfare economics, long-time occupant of the Chair of Political Economy at Cambridge University (1908–43), and author of hundreds of articles, pamphlets and books.

As Alfred Marshall's successor, he embraced, refined and extended the analytical framework that his master had painstakingly constructed. He also lived long enough to witness its disintegration at the hands of a generation of economists who had lost their tolerance for its limitations.

Life and Career

A.C. Pigou was born on 18 November 1877 at Ryde, Isle of Wight, England, and died in Cambridge on 7 March 1959. He attended Harrow (1891–96), emerging as a brilliant scholar and athlete who harboured a shyness of women that bordered on panic. Contrary to common belief, Pigou was no misogynist. He advocated paid maternity leaves for factory workers, voted for women's degrees at Cambridge University, and played a decisive role in creating a lectureship for the young Joan Robinson.

Pigou entered King's College, Cambridge on a Minor Scholarship in History and Modern Languages (1896). However, his interests spanned poetry, moral philosophy, politics and economics. His achievements were stunning: a First in the undivided History Tripos (1899) and another in Part II of the Moral Sciences Tripos with special distinction in political economy (1899), the Chancellor's Medal for English Verse (1899), the Burney Prize (1901), the Cobden Prize (1901) for an essay that secured him a fellowship at King's (1902), the Adam Smith Prize (1903) for work that formed the basis of his Jevons Memorial Lectures at University College, London (1903–4), and a Girdlers' lectureship (1904) that he held until his election to the Chair of Political Economy (30 May 1908).

Although significantly influenced by Henry Sidgwick, Pigou was the foremost disciple of Alfred Marshall, who was impressed by his protégé on several grounds. Pigou's 'exceptional genius', evident in his masterful thesis, foretold a future as 'one of the leading economists of the world'. He knew the proper role of economic theory: an instrument for social betterment, not intellectual gymnastics. Pigou fought for Marshall's brainchild, the independent Economics Tripos (established in 1903), and personally funded lectureships, prizes and book acquisitions. He shared Marshall's commitment to free trade, using his publications (Pigou 1904, 1906) and superb oratory skills – honed at the Cambridge Union Society of which he was President (1900) – to promote it. Together with Marshall, he signed the notorious Economists' Manifesto that rejected the Tariff Reform Proposal (1903) of Joseph Chamberlain. It is not surprising that Marshall's face beamed with delight when Pigou was chosen as his successor. He had manipulated the election in favour of the 30 year-old Pigou, embittering his old friend H.S. Foxwell, a serious contender.

Pigou's *Wealth and Welfare* (1912) – a synthesis of Marshall's engine of analysis and Sidgwick's categories of market failure and imperfections – laid the foundation for *Economics of Welfare* (1920), *Industrial Fluctuations* (1927a), and *A Study in Public Finance* (1928b). Taken together, these books covered most of the territory of general economics. *Industrial Fluctuations* was later complemented by *The Theory of Unemployment* (1933), which received a harsh and sophistical critique at the hands of J.M. Keynes in *The General Theory of Employment, Interest, and Money* (1936). Although faithful to the classical doctrine, *Employment and Equilibrium* (1941) – arguably the first textbook in macroeconomics – employed an IS-LM version of *The General Theory* and offered a careful analysis of the differences between Keynesian and classical economics. Pigou's other works included *Unemployment* (1913), *The Political Economy of War* (1921), *The Economics of Stationary States* (1935), many collections of essays, as well as books and pamphlets that he characterized as 'low-brow', among them the highly successful *Socialism versus Capitalism* (1937b), *Lapses from Full Employment* (1945) and *Income: An Introduction to Economics* (1946). The rise of a Cambridge School of economics was in large measure due to Pigou's articulation of Marshall's organon (see, for example, Pigou's classic exposition of Cambridge monetary theory, 1917). Generations of economists – among them Dennis Robertson, Joan and Austin Robinson, and Richard Kahn – learned Marshall in Pigou's lectures, which were legendary for their clarity and logical rigour.

Pigou was not a public man. His aversion to discussions of economics outside 'the home' extended to a distaste for conferences. Acting on his sense of public obligation, he served on several government committees – among them the Chamberlain Committee on the Currency and Bank of England Note Issues (1924–5), which recommended a return of sterling to its pre-war level, imposing immense costs on British labour. Disillusioned by British economic policies in the 1930s, he withdrew from public life, making only occasional ritually obligatory appearances before commissions.

Pigou's personal life also became increasingly hermetic. By the 1940s, the high-spirited, companionable young man of the Edwardian era was regarded as a recluse. As a conscientious objector, he never recovered from the experience of the carnage of the Great War, which he observed first-hand as a driver in the Friends' Ambulance Unit, commanded by his student and friend Philip Noel-Baker. Beginning in the mid-1920s, severe cardiac fibrillation (irregular heartbeat) curtailed his mountaineering – he was a deft climber introduced to the sport by the economic historian J.H. Clapham. This condition left him permanently anxious over his health. Finally, Pigou watched with dismay as the Keynesian Revolution destroyed the Edwardian intellectual culture of high civility in Cambridge economics. In time, he rose above his own angry response to Keynes's gratuitous depiction of classical economists as 'a gang of incompetent bunglers' (Pigou 1936, p. 115). But as relations between Keynes's disciples and Dennis Robertson became increasingly hostile, he grew more remote and diffident. In his judgement, Joan Robinson's dogmatic instruction of Keynesian economics turned undergraduates into 'identical sausages', and under Keynes's stewardship in the 1930s the *Economic Journal* violated its mission of representing different schools of thought 'with equal impartiality'.

Theoretical Contributions

Pigouvian economics is grounded in utilitarian moral philosophy: creating the greatest good –

Pigou's cognate of welfare – for the greatest number of people. Its analysis is limited to economic welfare: satisfactions that, directly or indirectly, can be related to the measuring rod of money. Up to a point, money, which measures the intensity of desires, performs well as a proxy for satisfaction. However, the human 'telescopic faculty' irrationally discounts future satisfactions, resulting in inadequate savings, insufficient investment in tunnels or forests, depletion of natural resources, and extinction of animal species. Pigou assumes that, as a rule, economic and total welfare are positively related. Anticipating contemporary research on happiness, he also recognizes the importance of factors that contribute to non-economic welfare such as relative status, social capital, political freedom, and moral quality of life.

Economic welfare may improve if its objective counterpart, the national product, is increased in size, distributed more evenly, and made more stable.

Optimal Resource Allocation

Integrating Marshall's marginal analysis and Sidgwick's distinction between private and public interests, Pigou produces some of the most important concepts (1912) and diagrams (1910) of welfare economics: marginal private and social net products (benefits and costs in contemporary parlance). In the absence of 'costs of movement' – associated with geographic and occupational reallocation of resources – the allocation of resources by competitive markets achieves universally equal marginal private net products. However, the production of ideal output requires equality of marginal social net products. Where private and social net products diverge, there is a prima facie case for reallocation of resources (1932, p. 136).

In Pigou's competitive economy, social and private benefits diverge in three different respects. First, a principal–agent problem arises when owners of land contract out its use to tenants. Since some benefits of the agent's investment accrue to the principal on termination of the contract, investment levels are not socially optimal. Pigou's remedies are limited to modifying contractual specifications between the two parties, presumably because low transactions costs render government action unnecessary.

Second, economic transactions between two agents may render incidental services or disservices to third parties, who cannot be forced to pay for the benefits or compensated for the costs. Unlike contemporary economists, Pigou does not distinguish public goods and externalities. Positive spillovers are a combination of public goods and beneficial externalities: lighthouses that benefit free-riding ships; private parks and forests that improve air quality; roads and tramways that improve the value of neighbouring land; privately owned lamps that shed light on streets; items of smoke-prevention equipment that benefit buildings, vegetables, clothes, and air quality; and 'most important of all' scientific research that leads to inventions, innovations and 'discoveries of high practical utility' (1932, p. 185). Negative spillovers are harmful externalities: a landlord raises rabbits that overrun a neighbour's property; a firm builds a factory in a densely populated area, destroying its amenities and injuring family health and productivity; automobile operators drive cars that wear out the surface of roads; and producers sell alcoholic beverages that increase crime. The 'crowning illustration' of negative externalities is women's factory work, especially immediately before and after childbirth, which damages the health of the fetus and increases infant mortality (1932, pp. 185–7).

Since it is difficult to internalize positive or negative externalities through contractual modifications, the state may offer 'extraordinary encouragements' or 'extraordinary restraints' as remedies, most obviously taxes and 'bounties'. In Pigou's era, a variety of taxes had already been imposed on alcoholic beverages, roads, gasoline and car licences. Bounties ranged from complete government provision (police protection and cleaning slums) to grants for scientific research. Pigouvian solutions went beyond taxes and subsidies to include patent enforcement, provision of information and training, and paid maternity leaves. In cases such as urban planning, where 'the inter-relations of the various private persons affected [are] highly complex', the state may have to exercise 'authoritative control' because the invisible hand fails to 'tackle the collective problems of beauty, of air and of light' (1932, pp. 193–6; also see 1947, pp. 94–100).

Careful readers of Pigou will note that much of Ronald Coase's critique of his analysis (Coase 1960) is misplaced. Pigou stressed that on issues of policy he always spoke with an 'uncertain voice' (Pigou 1932, p. 10), carefully considering the costs and benefits of proposed solutions. Government action entails allocative, administrative and political costs. Redeployment of labour, land and capital is also costly. It follows that the goal of achieving ideal output should be subjected to a cost–benefit analysis that shows 'at which point the advantage of getting closer is outweighed by the complications, inconvenience and expense involved in doing so' (Pigou 1932, p. 315).

Third, in his early work (1912), Pigou argued that private and social benefits diverge if industries exhibit increasing or decreasing costs. Under decreasing returns, a small increase in the output of one firm creates external diseconomies for the industry by increasing the price of fixed factors. Under increasing returns, a small rise in the output of one firm creates external economies for the industry. A prima facie case could therefore be made for taxing increasing-cost and subsidizing decreasing-cost industries. Pigou's critics – Allyn Young and Dennis Robertson – pointed out that the two types of returns are essentially different phenomena: external economies – technological change and managerial breakthroughs – are irreversible social gains. External diseconomies – increased factor prices – are not social costs since they merely transfer purchasing power from producers to factor owners. The second edition of *The Economics of Welfare* (1924) conceded this point, with the proviso that foreign owners do not capture the increased rents.

In 1926, Piero Sraffa argued that increasing and decreasing returns are incompatible with Marshall's competitive, partial-equilibrium assumptions. Under increasing costs, for instance, a marginal increase in the output of a firm in a given industry increases the price of fixed factors for all industries that use them. Relative prices may change as a result, rendering Marshallian assumptions logically incoherent since industry supply and demand become interdependent. Although economies and diseconomies that are external to the firm but internal to the industry do not generate the same logical problem, they are rare empirically. Pigou (1927b) concluded that, although increasing costs were incompatible with his framework, he could not logically rule out external economies. In 1928, he published the standard textbook analysis of stable equilibrium in a competitive firm (1928b). The costs of the equilibrium firm (a theoretical entity based on Marshall's representative firm) are a function of its own output and that of the industry. Although the industry may experience increasing or constant returns, the equilibrium firm is always at equilibrium when industry price is equal to its marginal and (the minimum of) average costs. U-shaped average and marginal cost curves for the equilibrium firm complemented the mathematical treatment, perhaps the first time that such diagrams were published in English. External economies shift the equilibrium firm's cost curves.

As a rule, monopolistic conditions create discrepancies between private and social benefits. Pigou argues that their implications for welfare must be evaluated on a case-by-case basis. The incidence of discrepancies depends on whether a monopoly practices price discrimination of the first, second or third degree. State control and state operation of natural monopolies have different ramifications for welfare. Oligopolistic market structures, however, create unequivocal social costs irrespective of output: wasteful advertising, exploitation of workers – defined as payment below the value of marginal product – customer deception, reduction of upward mobility by forcing small entrepreneurs out of the market, constraints on inventions and innovations, and Tayloristic practices that dull worker initiative. Pigouvian remedies range from taxes and prohibitions to encouragement of small business.

Income Redistribution

Redistribution schemes that favour the poor but leave the national product intact are likely to improve economic welfare. However, both the expectation and the fact of such transfers may produce disincentives that reduce the national product. The implication is not inaction. Rather, the state should design redistributive measures based on a comprehensive knowledge of legal, psychological and institutional factors. If capital is subject to double taxation, its flight is less probable. If economic actors target a specific level of savings, inheritance taxes may not affect investment activity. If redistributed income is used to train workers with uncommon abilities, its rate of return may surpass the return on investment in physical capital. Finally, taxation may not discourage the rich if it leaves their relative income intact. Pigou's theoretical analysis of interdependent utility (welfare) – based on reference groups, relative income, snob and bandwagon effects – anticipates Duesenberry's and Leibenstein's by some 45 years (Pigou 1903).

Transferring one dollar from the rich to the poor increases economic welfare because 'it enables more intense wants to be satisfied at the expense of less intense wants' (Pigou 1932, p. 89). This proposition assumes that representative members of different income groups have equal capacities for satisfaction. In 1932, Lionel Robbins claimed that such interpersonal comparisons are normative judgments and have no place in science. The ensuing attempts to establish a positivist welfare economics engaged such luminaries as Hicks, Kaldor, Scitovsky, Little, Bergson and Arrow. The results produced a sophisticated theoretical apparatus but confirmed Pigou's belated response to Robbins that without such comparisons every 'apparatus of practical thought' will collapse (Pigou 1951, p. 292). In recent decades, the recognition that all sciences make normative claims has become received wisdom in the philosophy of science. With the demise of doctrinaire positivism, economists seem more willing to venture into the territory of interpersonal comparisons, as contemporary happiness research suggests. This research provides new grounds for reconsidering the unexploited resources of Pigouvian welfare economics.

Industrial Fluctuations and Unemployment

Long spells of unemployment have serious deleterious effects – malnutrition, permanent damage to the capabilities of youth, loss of skills and work ethic, alcoholism, a 'haunting' sense of insecurity and uncertainty, and the destruction of self-respect and self-confidence – that cannot be reversed in good times. Thus a prima facie case for macroeconomic stability is evident.

Pigou's theory of unemployment can be elucidated by using the language of supply and demand. Aggregate labour supply is vertical, even though individual labour supply curves may be upward sloping or backward bending. Aggregate labour demand – difficult to construct due to sectoral interdependence – is downward sloping and dependent on marginal product. Since unemployment is always positive, it can be explained only by movements in wages and the demand for labour.

Pigou distinguishes two types of unemployment. Short-run involuntary unemployment – a term he may have coined in 1913 – occurs because of frequent changes in labour demand and real wages. Although prices vary, real wages fluctuate because nominal wages remain sticky. (a) A perpetually flexible nominal wage is impracticable due to high administrative costs, which become more significant if 'elaborate and formal arbitration proceedings' are instituted to resolve capital-labour conflicts (Pigou 1913, pp. 92–3). (b) Some wage rigidity is preferred: while workers want stable living standards, employers are obliged to deliver products at prices previously negotiated. (c) The duration of recessions and recoveries is unpredictable; it is not worthwhile to alter wages if the state of the economy is ephemeral. (d) Due to mutual mistrust, workers and firms alike resist wage changes, fearing that they may be irreversible. (e) Employees and employers suffer from money illusion, the latter resisting wage increases and the former refusing wage cuts.

Contrary to Keynes's straw-man depiction, Pigouvian labour demand fluctuates due to general and wave-like swings in expectations of profits. Three sets of factors affect expectations: real causes such as crop size or technological breakthroughs; monetary variables, which are restricted to exogenous shifts in credit under the gold standard; and psychological factors, which occur spontaneously or as a consequence of the other two variables. Undue pessimism or optimism may be magnified because psychology, output, and debt–credit linkages create sectoral interdependence.

The amplitude of business cycles depends on the institutional structure of the economy: monetary policy, the pricing strategy of firms, income maintenance programmes, wage policy and unions. Although limited by the quality and quantity of data at his disposal, Pigou tries to quantify factors that cause business cycles or affect their amplitude. Removing monetary or psychological factors would each reduce the amplitude by one-half, crop variation by one-quarter, wage rigidity by one-eighth and price rigidity by one-sixteenth. It is clear that Pigou does not regard high real wages as the single or even the most important cause of short-run unemployment. In many cases, high wages and unemployment are both effects of factors such as 'bursting of a gigantic bubble of unwarranted optimism, with a heavy fall in price' (Pigou 1929, pp. 200–1). Short-run unemployment can be reduced proactively through distribution of information, price stability, or interest-rate manipulation. Reactive policies that dampen the impact of unemployment range from (public work projects to guarantees of interest or subsidies for employers. Although Pigou favours wage flexibility at the theoretical level, he does not consider it a viable political option.

Pigou analyses long-run unemployment on stationary-state assumptions, ruling out changes in expectations, tastes, net investment, productivity, and technology. The only conceivable unemployment under these conditions is an 'intractable minimum' that resembles the natural rate of unemployment. It is caused by frictions, immobility, public opinion, the practical impossibility of setting wages according to marginal productivity, and unions. Collective bargaining introduces indeterminacy, which he analyses in a quasi-game-theoretic framework (Pigou 1905). Employers and employees negotiate money wages within a 'range of indeterminateness'. The upper limit depends on unions' reluctance to demand a wage so high that it would result in layoffs. The lower limit is determined by employers' recognition that a wage that reduces the available supply of labour is too low. Peaceful wage bargains are conducted within a narrower range determined by the 'sticking point' of each party: a certain minimum below which workers would rather strike and a maximum above which employers would prefer shutdowns.

Firms often have bargaining power to exploit workers but may choose not to, recognizing that low wages affect the productivity of workers they want to retain for the long period. This results in unemployment in a casual labour market. The magnitude of joblessness is determined by a Harris–Todaro comparison of the expected wage – 'the wage-rate multiplied by the chance of employment' (Pigou 1913, p. 55) – with wages elsewhere. Unemployment is not an inevitable outcome if outsiders (low-wage workers) know that insiders (high-wage workers) are irreplaceable.

To reduce long-run unemployment, the state may attempt to educate the unskilled and try to improve wage flexibility. The effective demand ramification of wage flexibility was a major point of contention between Keynes and Pigou (see Pigou 1937a; Keynes 1937). Although Pigou was finally persuaded by Kaldor (1937) to take such effects into account (Pigou 1938, 1941), he discounted them based on the well-known Pigou effect: lower money incomes and prices would increase the value of real balances, reducing and ultimately eliminating the individual's desires to save out of any assigned real income (Pigou 1943, p. 349). In Pigou's opinion, Keynes's true contribution was not substantive but analytical: no one before him had constructed a model of the aggregate economy that incorporated both real and monetary factors. But Pigou (1950) also maintained that Keynes's analytical framework was too limited to be suitable for direct practical application.

Legacy

Economists have generally judged Pigou's work on Robbinsian, Keynesian or Coasean premises, ignoring his important contributions to the theories of value, distribution, business cycles, public finance, index numbers, and evaluation of real national income. Pigou's neglected contributions to labour economics, which anticipate Hicks's work by a quarter of a century, are especially noteworthy. *Wealth and Welfare,* hailed by Schumpeter as 'the greatest venture in labor economics ever undertaken by a man who was primarily a theorist' (1954, p. 948), and his numerous other works on labour and unemployment demonstrate an acute understanding of the importance of a remarkable range of phenomena explored by subsequent economists – implicit contracts, internal labour markets, labour market segmentation, wage rigidity, human capital theory, and collective bargaining. Alfred North Whitehead famously held that 'a science which hesitates to forget its founder is lost'. Economists have not found it difficult to forget the founder of welfare economics, with regrettable consequences that Whitehead did not envision.

See Also

- ▶ Financial Intermediation
- ▶ Interpersonal Utility Comparisons
- ▶ Keynesian Revolution
- ▶ Labour Economics
- ▶ Unemployment
- ▶ Welfare Economics

Selected Works

1903. Some remarks on utility. *Economic Journal* 13: 58–68.

1904. *The riddle of the tariff.* London: R. Brimley Johnson.

1905. *Principles and methods of industrial peace.* London: Macmillan.

1906. *Protective and preferential import duties.* London: Macmillan.

1910. Producers' and consumers' surplus. *Economic Journal* 20: 358–370.

1912. *Wealth and welfare.* London: Macmillan.

1913. *Unemployment.* London: William and Norgate.

1917. The value of money. *Quarterly Journal of Economics* 32: 38–65.

1920. *The economics of welfare.* London: Macmillan.

1921. *The political economy of war.* London: Macmillan.

1927a. *Industrial fluctuations.* London: Macmillan.

1927b. The laws of diminishing and increasing costs. *Economic Journal* 37: 188–197.

1928a. An analysis of supply. *Economic Journal* 38: 238–257.

1928b. *A study in public finance.* London: Macmillan.

1929. *Industrial fluctuations,* 2nd edn. London: Macmillan.

1932. *The economics of welfare,* 4th edn. London: Macmillan.

1933. *The theory of unemployment.* London: Macmillan.

1935. *The economics of stationary states.* London: Macmillan.

1936. Keynes' J.M. *General theory of employment, interest and money. Economica* 3: 115–132.

1937a. Real wages and money wage rates in relation to unemployment. *Economic Journal* 47: 405–422.

1937b. *Socialism versus capitalism.* London: Macmillan.

1938. Money wages in relation to unemployment. *Economic Journal* 48: 134–138.

1941. *Employment and equilibrium.* London: Macmillan.

1943. The classical stationary state. *Economic Journal* 53: 343–351.

1945. *Lapses from full employment.* London: Macmillan.

1946. *Income: An introduction to economics.* London: Macmillan.

1947. *A study in public finance,* 3rd edn. London: Macmillan.

1950. *Keynes's general theory: A retrospective view.* London: Macmillan.

1951. Some aspects of welfare economics. *American Economic Review* 41: 287–302.

Bibliography

Aslanbeigui, N. 1990. On the demise of Pigouvian economics. *Southern Economic Journal* 56: 616–627.

Aslanbeigui, N. 1992. Pigou's inconsistencies or Keynes's misconceptions? *History of Political Economy* 24: 413–433.

Aslanbeigui, N. 1996. The cost controversy: Pigouvian economics in disequilibrium. *European Journal of the History of Economic Thought* 3: 275–295.

Aslanbeigui, N. 1998. Unemployment through the eyes of a classic. In *Keynes and the classics reconsidered*, ed. J.C.W. Ahiakpor. Boston: Kluwer.

Aslanbeigui, N., and S.G. Medema. 1998. Beyond the dark clouds: Pigou and Coase on social cost. *History of Political Economy* 30: 601–625.

Aslanbeigui, N., and G. Oakes. 2007. The editor as scientific revolutionary: Keynes, *The Economic Journal,* and the Pigou affair, 1936–1938. *Journal of the History of Economic Thought* 29: 1–34.

Coase, R.H. 1960. The problem of social cost. *Journal of Law and Economics* 3: 1–44.

Collard, D. 1981. A.C. Pigou, 1877–1959. In *Pioneers of modern economics in Britain*, ed. D.P. O'Brien and J.R. Presely. Totowa: Barnes and Noble Books.

Collard, D. 1996. Pigou and modern business cycle theory. *Economic Journal* 106: 912–924.

Cooter, R., and P. Rappoport. 1984. Were the ordinalists wrong about welfare economics? *Journal of Economic Literature* 22: 507–530.

Kaldor, N. 1937. Prof. Pigou on money wages in relation to unemployment. *Economic Journal* 47: 745–753.

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

Keynes, J.M. 1937. Prof. Pigou on money wages in relation to unemployment. *Economic Journal* 47: 743–745.

O'Donnell, M.G. 1979. Pigou: An extension of Sidgwickian thought. *History of Political Economy* 11: 588–605.

Robbins, L. 1932. *An essay on the nature and significance of economic science*. London: Macmillan.

Robertson, D.H. 1924. Those empty economic boxes: A rejoinder. *Economic Journal* 34: 16–30.

Saltmarsh, J., and P. Wilkinson. 1960. *Arthur Cecil Pigou, 1877–1959*. Cambridge: Cambridge University Press.

Schumpeter, J.A. 1954. *History of economic analysis*. New York: Oxford University Press.

Shiller, R.J. Ultimate sources of aggregate variability. *American Economic Review* 77: 87–92.

Solow, R.M. 1980. On theories of unemployment. *American Economic Review* 70: 1–11.

Sraffa, P. 1926. The laws of returns under competitive conditions. *Economic Journal* 36: 535–550.

Young, A. 1913. Pigou's wealth and welfare. *Quarterly Journal of Economics* 27: 672–686.

Pigouvian Taxes

Agnar Sandmo

Abstract

Pigouvian taxes are taxes designed to correct for negative external effects. The idea is originally due to Pigou (1920), and has received increased attention in recent years because of the concern with environmental issues. This article sets out the basic theoretical argument and considers the modifications of the theory that have to be made when these taxes are seen in the context of an otherwise distortionary tax system. It also briefly considers the issue of the 'double dividend' from a green tax reform.

Keywords

Distortionary tax; Double dividend; Externalities; Lump sum taxes; Marginal cost of public funds; Optimal taxation; Partial equilibrium; Payroll tax; Pigou, A. C.; Pigouvian taxes; Ramsey tax; Substitutes and complements; Tax base; Tax wedge

JEL Classifications

H2

'Pigouvian taxes' is the generic term for taxes designed to correct inefficiencies of the price system that are due to negative external effects. In partial equilibrium terms, the basic idea can be presented as follows: under competitive conditions, utility-maximizing consumers will equate their marginal benefit to the market price Q; we may write this as $MB = Q$. Similarly, profit-maximizing producers will set their marginal private cost equal to the price, so that $MPC = Q$. In the absence of externalities, marginal private and social costs coincide: $MPC = MSC$. Consequently, market equilibrium implies that $MB = MSC$, which is the condition for efficient resource allocation. If there are negative external effects related to the production or consumption of the good in question, the marginal social cost is higher than the marginal private cost: $MSC>MPC$. If the market prices facing producers and consumers are identical, this implies that $MB < MSC$. To restore efficiency, we may levy a tax on the commodity, so that the consumer price is Q while the producer price is $Q - t$. In the new equilibrium we have that $MB = Q$ and $MPC = Q - t$; it follows that $MB = MPC + t$. Since we wish the equilibrium to satisfy the condition that $MB = MSC$, we must have $t = MSC - MPC$, which we may define as the marginal social damage. Accordingly, the optimal Pigouvian tax internalizes the externality; producers act as if they took account of the marginal social damage associated with the production of the commodity.

This idea was first expressed by Pigou, especially in his *Economics of Welfare* (1920). He mentions a number of examples of what he calls divergence between 'social and private net product', for example, production activities generating smoke from factory chimneys that create adverse consequences for consumers in the form of damage to buildings, increased expenses for washing clothes, house-cleaning and indoor lighting. These inefficiencies can be corrected, he says, by 'imposing appropriate rates of tax on resources that tend to be pushed too far'; he also points out that cases of positive externalities where $MSC < MPC$ can be corrected by means of subsidies or 'bounties' (Pigou 1920; 1932, p. 184). In his later book, *A Study in Public Finance*, he claims that

> [there] will necessarily exist a certain determinate scheme of taxes and bounties, which, in given conditions, distributional considerations being ignored, would lead to the *optimum* result. (Pigou 1928; 1947, p. 99)

An interesting and important question concerns the choice of the tax base. On what should the Pigouvian tax be levied? From a theoretical point of view, the correct tax base is the one that affects the crucial margin of decision. In the factory smoke example, the best tax base is actually the amount of smoke emission. A tax on coal is an imperfect instrument to the extent that it also affects margins that are irrelevant for smoke emission, and this is even more true for a tax on the output produced by the factory. Some would therefore reserve the term 'Pigouvian tax' for the tax on smoke emission, but in the literature it has become common to use the concept to refer to all cases where the policy motivation is to correct for negative externalities.

For a long time, Pigouvian taxes led an obscure life in the public economics literature; thus, in the famous treatise by Musgrave (1959), the subject is barely mentioned. However, with the increased concern for the environment that rapidly gained ground from the late 1960s, economists became much more interested in this form of tax policy both as a tool for environmental policy and as an efficient source of revenue for the public sector.

Distortionary Taxes

The partial equilibrium approach is based on some simplifying assumptions. First, it focuses solely on the market for the 'commodity' (final good, factor of production or emission) that gives rise to the externality, while neglecting the interconnections with other markets. Second, it assumes, rather implicitly, that there are no other violations of the efficiency conditions in the economy, so that the design of Pigouvian taxes does not need to take into account the presence of other distortions. Third, as also emphasized by Pigou, it ignores distributional concerns.

All these simplifications must be overcome if one wishes to analyse Pigouvian tax policy within the context of the overall tax system. There is actually one tax system in which the partial equilibrium analysis is valid, and that is the assumption that the rest of the requirement for public sector revenue can be satisfied by means of individualized lump sum taxes. This leads to a 'first-best' allocation: tax revenue is raised without distortions of the price mechanism, and the desired income distribution can be achieved without loss of efficiency. The only commodity taxes that are used are the Pigouvian taxes on commodities that generate negative external effects. But lump sum taxes are not policy instruments that can be used realistically. Instead, governments have to rely on direct and indirect taxes, and these will create tax wedges and distortions of private incentives. What is the role of Pigouvian taxes within the context of an otherwise distortionary tax system?

One might perhaps come to think that in such a setting Pigouvian considerations should affect the taxes on all goods: for example, there might be a case for subsidizing substitutes and taxing complements to the harmful commodities. However, it was shown in Sandmo (1975) that in an optimal system of commodity taxes the integration of Pigouvian taxes with the Ramsey (1927) objective – minimizing efficiency loss for a given tax revenue – takes a strikingly simple form. If there is one commodity that creates a negative externality, the tax on this commodity can be expressed as a weighted average of Ramsey and Pigou terms, while other taxes contain only a Ramsey term. Formally, suppose that there are a number of taxed goods ($i = 1,..., n$) and that the externality is generated by the nth good. Suppose for simplicity that all cross-elasticities between the taxed goods are zero, so that Ramsey taxes can be characterized by the inverse elasticity formula.

Then the optimal tax system can be written as follows:

$$t_i = \alpha(-1/\varepsilon_i)(i = 1,...,n-1).$$

$$t_n = \alpha(-1/\varepsilon_n) + (1-\alpha)\delta_n.$$

Here ε_i is the own price elasticity of commodity i, and δ_n is the marginal social damage of commodity n. α is a parameter that characterizes the tightness of the government budget constraint. If the budget is extremely tight, all weight is on the need for revenue. Then $\alpha = 1$, the tax rates are chosen so as to maximize revenue, and Pigouvian taxes play no role in the tax structure. However, in the happy situation where the revenue from Pigouvian taxes is exactly sufficient to meet the government's revenue requirement, $\alpha = 0$, and no other taxes are desirable. It can be shown that the 'additivity' property of the optimal tax system continues to hold when distributional considerations are incorporated into the model, but in that case the weights on the inverse elasticity and the marginal social damage will have to reflect distributional concerns in addition to those of efficiency.

The Double Dividend and the Marginal Cost of Funds

In recent years there has grown up a strong interest in 'green tax reforms'. Such reforms would reduce conventional distortionary tax rates and compensate for the loss of tax revenue by introducing more Pigouvian taxes. A popular view of the gain from this kind of reform is that society would reap a 'double dividend'. First, higher Pigouvian taxes would create an improved environment; second, lower distortionary taxes would imply a more efficient tax system. This argument has a strong appeal to economic intuition; however, as often happens, when one comes to study it more closely, it turns out to contain some complicating elements. The crucial point to note is that the effects of Pigouvian taxes interact with those of the distortionary taxes. If for example the existing tax system has a high marginal tax rate on labour income, an increase of Pigouvian taxes together with a lowering of other indirect tax rates might exacerbate the labour market distortion if the externality-creating goods are complementary with labour supply. This argument does not imply that the argument in favour of the double dividend is groundless. It simply means that one has to be careful in taking account of the interaction between markets for taxed goods before predicting a double dividend.

Another version of the double dividend argument focuses on unemployment. If the basic cause of unemployment is that employers' labour cost is above the market-clearing wage, a promising tax reform might be to reduce the payroll tax while increasing Pigouvian taxes. The double dividend in this case would be a better environment and lower unemployment. Again, the consensus of professional opinion seems to be that this is indeed a possible outcome, but it is by no means assured. For example, in a unionized labour market much will depend on the combined incidence of a reduced payroll tax and higher indirect taxes on union wage demands. For further discussion of both versions of the double dividend, see Bovenberg (1999) and Sandmo (2000).

Related to the question of the double dividend is the relationship between Pigouvian taxes and the marginal cost of public funds (MCF), a concept whose origin can also be traced to Pigou: see Atkinson and Stern (1974). With distortionary tax finance, the direct resource cost of public goods should be multiplied with an MCF adjustment factor which exceeds one. Since Pigouvian taxes actually increase the efficiency of the market mechanism, one might expect that for this type of tax finance one would have MCF < 1. Theoretical analysis has shown that this is indeed likely to be true in a number of cases, but that here too one needs to pay attention to the interaction of distortionary and Pigouvian taxes.

See Also

- ▶ Environmental Economics
- ▶ Optimal Taxation

Bibliography

Atkinson, A.B.., and N. Stern. 1974. Pigou, taxation and public goods. *Review of Economic Studies* 41: 119–128.

Bovenberg, A.L. 1999. Green tax reforms and the double dividend: An updated reader's guide. *International Tax and Public Finance* 6: 421–443.

Musgrave, R.A. 1959. *The theory of public finance*. New York: McGraw-Hill.

Pigou, A.C. 1920. *The economics of welfare*. 4th ed, 1932. London: Macmillan.

Pigou, A.C. 1928. *A study in public finance*. 3rd ed, 1947. London: Macmillan.

Ramsey, F.P. 1927. A contribution to the theory of taxation. *Economic Journal* 37: 47–61.

Sandmo, A. 1975. Optimal taxation in the presence of externalities. *The Swedish Journal of Economics* 77: 86–98.

Sandmo, A. 2000. *The public economics of the environment*. Oxford: Oxford University Press.

Pirou, Gaetan (1886–1946)

Roger Dehem

Born at Le Mans in 1886; died in Paris, 1946. A doctor in law (Rennes, 1909) and economics (Paris, 1910), Pirou taught at the Institut Français in London (1913–14), at Rennes, Milan, Algiers and Bordeaux (1920–26) before his appointment to the chair of history of economic thought in Paris (1927). He later taught at the Ecole des sciences politiques (1940–46) and was editor of the *Revue d'économie politique* (1935–46).

Pirou was foremost a distinguished teacher and a keen historian of economic and social thought. With a traditional middle-class and legal background, he first searched for idealistic truths on the left of the intellectual spectrum, with studies on Proudhon and Sorel. He then wrote *Les doctrines économiques en France depuis 1870* (1925). The assessment of ideologies, systems and policy experiments in the 1930s was the subject of several essays. In economic theory, Pirou consistently adhered to the neoclassical paradigm. He contributed a comprehensive though non-mathematical survey of the Austrian and general equilibrium theories. His most original contribution can be found in a methodological introduction to political economy (1939).

Pirou is duly respected for his intellectual honesty, his balanced judgement, his search for objective truth beyond class prejudices. A constant concern of his was the contradiction between the power of scientific reasoning (Comte) and the persistence of irrationality (Sorel). He was an erudite scholar and an academic *par excellence*.

His work is to be appreciated as a pedagogic instrument rather than as a guide to action. Though eclectic, his thought was not without depth nor consistency.

Selected Works

1910. *Proudhonisme et syndicalisme révolutionnaire*. Paris: Rousseau.
1925. *Les doctrines économiques en France depuis 1870*. Paris: A. Colin.
1927. *Georges Sorel*. Paris: Marcel Rivière.
1929. *Doctrines sociales et science économique*. Paris: Sirey.
1932. *L'utilité marginale: de C. Menger à J.B.* Clark. Paris: Domat-Montchrestien.
1934a. *Les théories de l'équilibre économique: L. Walras et V. Pareto*. Paris: Domat-Montchrestien.
1934b. *La crise du capitalisme*. Paris: Sirey.
1938. *Essais sur le corporatisme*. Paris: Sirey.
1939a. *Introduction à l'étude de l'économie politique*. Paris: Sirey.
1939b. *Néo-libéralisme, néo-socialisme, néo-corporatisme*. Paris: Gallimard.

Pissarides, Christopher (Born 1948)

Richard Jackman

Abstract

Professor Christopher Pissarides was awarded the 2010 Nobel Prize in Economics jointly with Peter Diamond and Dale Mortensen 'for their analysis of markets with search frictions'. Though Pissarides is best known for his work in this area, it is only part of a very extensive research agenda which has covered numerous topics in theoretical and applied macroeconomics, and in particular in the analysis of labour markets. Within search theory his main contributions have been in integrating search theory into models of economic equilibrium and in examining its implications for efficiency and labour market policy.

Keywords

Labour markets; Macroeconomics; Search theory; Unemployment; Wage determination

JEL Classifications

B31; D83; E24; J21; J23; J63; J64

A native of Cyprus, Pissarides came to Britain in 1967 to study economics at the University of Essex. Essex was at that time a new university which had ambitions, amongst other things, to build up a top-class economics department with a strong emphasis on theory. On the strength of a first-class undergraduate degree, followed by a Masters with distinction, Pissarides left Essex in 1971 to study for his PhD at LSE. His thesis, on 'Individual Behaviour in Markets with Imperfect Information' (supervised by Michio Morishima) was the starting point for the extensive exploration of this topic in articles spanning close on 40 years.

Following a short spell at the University of Southampton, Pissarides returned to LSE as a Lecturer in 1976 and he has remained at LSE ever since. Throughout his career he has been a central figure in the LSE Economics Department, teaching macroeconomics at both undergraduate and graduate level, including in the recently revamped PhD programme. He became a Professor in 1986 and was Convenor (Head of Department) from 1996 to 1999. He has also been a leading figure in the LSE Centre for Labour Economics (which became the Centre for Economic Performance in 1990), where he was Programme Director for Macroeconomics from 1999 to 2007.

The late 1960s and early 1970s were a time of great intellectual ferment in many fields of economics, most significantly for labour market analysis by the development of search theory. The idea of search as the behavioural principle underlying the workings of decentralized markets was established most notably by the 'Phelps volume' (Phelps et al. 1970) and especially the paper by

Armen Alchian (1970). The papers in the Phelps volume stressed the role of imperfect information in price and wage setting by firms and in the labour and product market decisions of workers and consumers. One of its authors, Chris Archibald, had been at the University of Essex, but it is fair to say these ideas established themselves rapidly in the teaching of macroeconomics across the world, for example in the LSE MSc course then taught by Michael Parkin.

In the theory of unemployment the search model was something of a paradigm shift. Previously, unemployment had been analyzed mainly in terms of essentially static concepts, such as the level of aggregate demand or structural imbalances, rather than in terms of the behaviour of unemployed workers. The search model argued that unemployed workers would not want to sign up for any job, but would need to make some estimate of the best job they could reasonably hope to get over some realistic time horizon. They would then reject job offers (or perhaps more plausibly not apply for jobs) paying less than this estimate, and remain unemployed until they received a sufficiently good offer. The new approach allowed a more rigorous analysis of numerous labour market issues – for example the impact of unemployment benefits (which can clearly affect the relative advantages of taking a job today as against continuing to search for a better job tomorrow). For a survey of the development of search theory and the contributions of the three 2010 Nobel Laureates to it, see Albrecht (2011).

While most research in search theory focused on the microeconomic decision making of individual workers, Pissarides's distinctive contribution has been to integrate search theory into complete, general equilibrium models of the economy. The microeconomic decision for unemployed workers was seen as one of whether or not to accept a job offer. This problem was formalized, somewhat artificially, by assuming that each unemployed worker received one job offer in each period. Workers were assumed to be aware of the general distribution of wages prevailing in the economy. Their problem was to decide their cut-off, or reservation, wage such that jobs offering lower wages would be rejected while those paying at least the reservation wage would be accepted. An important insight of this model is that workers would set their reservation wage on the basis of their expectations concerning the wage distribution they faced, which might or might not be the same as the actual wage distribution. In this way, imperfect information could play a critical role in accounting for fluctuations in unemployment. For example if the actual distribution of wages was higher than expected, unemployed workers would more quickly come across jobs offering wages above their reservation wage and hence more quickly find a job they would accept. This would lead to a faster outflow of people from unemployment and hence a decline in aggregate unemployment.

Pissarides made numerous contributions, both theoretical and empirical, to the study of the search behaviour of unemployed workers, considering amongst other things the effect of state employment agencies, job advertising, taxes and subsidies, and more recently the analysis of on-the-job search (Pissarides 1994). His distinctive contribution, though, has been the development of general equilibrium models incorporating search. This work appeared in a series of papers published during the 1980s, culminating in his classic text *Equilibrium Unemployment Theory* (Pissarides 1990, 2000).

The general equilibrium approach as developed by Pissarides highlighted the significance of job vacancies as the demand side counterpart to unemployment. Whilst, for example, the 1970 Phelps volume (cited above) makes almost no reference to job vacancies, in Pissarides's equilibrium model a key element is the matching function, which relates job hires (the outflow from unemployment) to unemployed workers and job vacancies. Having a ‘quantity’ variable on the demand side (in place of the assumption of job offers arriving at a constant rate) allows the search model to engage with key questions about variations in demand conditions in the labour market. Rather than receiving one job offer per period, the probability of a worker getting an offer depends on the ratio of vacancies to unemployment in the relevant labour market. The matching function, and the vacancy/unemployment ratio, took over from the reservation wage as the key concepts in the analysis of search unemployment.

The matching function was itself not invented by Pissarides but may none the less claim an LSE parentage. The relationship between unemployment and job vacancies has intrigued British economists for a long time, and the graphical representation of the relationship over time, in the form of the unemployment/ vacancy (u/v) curve was first charted by Beveridge (1944). The idea that the u/v curve might be generated from a matching function is due to another LSE author, Richard Lipsey (1960).

If vacancies are central to the theory it is clearly necessary to understand when firms choose to establish new jobs. Clearly this would depend not only on the return from having another worker in post, but also on the wage that the firm would pay and on the time it would take to fill the position. But wages and recruitment are of course linked, in that the higher the wage offer the easier it is to fill the vacancy. A complete model requires a theory of wage determination also. Given imperfections of information, both firm and worker have to ‘invest’ in a process of job search, and a successful job match must therefore offer a return adequate to compensate for this investment. The distribution of this return between firm and worker is determined by the wage. There are various approaches to how the wage gets determined, but, as may seem consistent with the decentralized approach of search theory, Pissarides assumed it would be the outcome of an individual bargain between firm and worker, and hence assumed the Nash bargaining solution. With these elements in place, it is possible to solve for the equilibrium of the system.

The comparative static properties of the equilibrium are fairly intuitive; in particular that unemployment will be higher if the replacement ratio (the ratio of unemployment benefits to the wage) is higher, because benefits support lengthier search; that for a given replacement ratio, unemployment will be unaffected by labour productivity; and that a more rapid rate of job destruction will increase unemployment.

Perhaps more interesting is the behaviour of the system out of equilibrium. For example, an increase in demand makes it profitable for firms to open up more vacancies, thus increasing the number of matches and hence the outflow from unemployment. As unemployment falls, firms find it more difficult to fill their vacancies and hence raise the wage they are prepared to offer. But higher wages make it less profitable to open up new jobs, so the number of vacancies falls back again. One of Pissarides’s best known papers (Pissarides 1985a) formalises this process. This paper is based on the assumptions that, while the rate of job separations (quits and lay-offs) is assumed exogenous, a firm can adjust its stock of vacancies immediately in response to changes in demand, and likewise that wages could also be adjusted instantly. The number of unemployed workers can, however, only change slowly. This is because the inflow into unemployment (job separations) is proceeding at a given exogenous rate, while the outflow from unemployment (job hirings) was also determined, albeit endogenously, by the stock of unemployed workers and of job vacancies. (Of course the rate of job hirings would increase immediately the number of vacancies increased, but it would still take time for the stock of unemployment to fall.) The model did not allow for the ‘sudden death’ of jobs that can arise where there are bankruptcies or firms suffer financial distress, and which would lead to a jump increase in unemployment.

To complete this analysis, then, involved a theory of job separations and in particular of jobs coming to an end, or job destruction. The key paper in this area was co-authored by Pissarides and fellow Laureate Dale Mortensen (1994). Within this model, a firm can close down jobs, just as it can open up vacancies, instantaneously in the face of some shock. This, however, introduces an asymmetry – a firm facing a positive shock can open up vacancies instantaneously and hence start recruiting, but hiring takes time, so unemployment falls only slowly. A firm facing a negative shock can close down jobs instantaneously, but this will lead to a sharp rise in unemployment.

The other main area within search theory on which Pissarides has worked is on the efficiency of the search equilibrium and the implications for labour market policy. The role of externalities in the search process and their impact on the efficiency of the search equilibrium were explored by fellow Laureate Peter Diamond in the early 1980s, and the issue was followed up in a number of papers by Pissarides, perhaps most notably in his paper on search intensity, job advertising and efficiency (Pissarides 1984). The idea here is that while more intensive job search involves a cost to the individual of more time and effort, it also offers the benefit of greater probability of finding work. But this benefit to the individual is not necessarily a benefit to society, since the job is no longer available for any other unemployed job-seekers. On the other hand, the more actively the unemployed search for work, the easier it will be for firms to recruit and hence the more willing they will be to open up vacancies. So there may be too much or too little job search, and labour market intervention may be needed to achieve something close to the optimal level.

In fact, of course, the intensity of search is already greatly affected by financial interventions, such as unemployment benefits, taxes and specific labour market inducements, such as subsidies for firms taking on particular types of workers. Most contentious in this area has been the implication of the search model that unemployment benefits reduce the incentive to search, to apply for jobs and to accept job offers. The design of unemployment compensation of course involves many other considerations (insurance against job loss, income support etc.) and has been the subject of innumerable studies. At a theoretical level Pissarides's most notable contribution was the analysis of the effects of benefits in a full equilibrium model (Pissarides 1985b), though given the offsetting externalities inherent in the search model it is unsurprising that the policy impact of this work has been quite limited.

Pissarides has made many other contributions to the theory of labour markets with search frictions, which have ranged from work on job-to-job movements to the integration of search to models of balanced growth. There is not space to describe this work in any detail. It may be noted though, that with the exception of the matching function, which is a quasi-macroeconomic type of relationship (essentially an empirical generalization based on aggregate data), Pissarides's work is very 'pure' -it assumes throughout that all agents behave rationally and all individual opportunities for profit are exploited. The *Equilibrium Unemployment Theory* book contains no facts, numbers or regressions. The equilibrium model abstracts from institutional factors, such as price rigidities or collective action, which in other models may lead to problems, though whether this is a strength or a weakness may be a matter of opinion.

Though not part of the Nobel citation, Pissarides also has an impressive range of empirical work. Clearly any macroeconomist working in the 1970s and 1980s cannot have ignored the global inflation in the 1970s nor the relapse into 'world-wide stagflation' in the early 1980s. At LSE, Pissarides was closely involved with the empirically orientated work being developed at the time at the Centre for Labour Economics (CLE) under the leadership of Richard Layard and Stephen Nickell. Even in the depths of the recession in the UK in the early 1980s, some CLE economists were struck by the observation that the number of job vacancies remained quite high, which they interpreted to mean that the high unemployment of that time was substantially an equilibrium phenomenon rather than the result of deficient demand. This provided an empirical counterpart to the significance of vacancies in Pissarides's theoretical work. Pissarides co-authored one of the key papers in this area with Richard Layard and Richard Jackman (Jackman et al. 1983). This was followed by a large number of papers by Pissarides and other CLE authors analysing the nature and causes of high unemployment and discussing policies to reduce the equilibrium (or natural) rate of unemployment.

Pissarides has made important empirical contributions to many issues in labour market policy.

These include structural issues such as regional imbalances and migration, skills shortages and technological change and questions of wage flexibility, perhaps most notably a paper co-authored with Dale Mortensen on technology shocks (Mortensen and Pissarides 1999). He has also written on employment protection and employment taxes and labour force participation, including female participation and, most recently, the impact of variability of hours. All these studies are characterized by a strong empirical grounding and a focus on policy, and in many cases are the product of joint work with other researchers at CLE/CEP.

Although it is not the main focus of his own work, Pissarides has also thrown his weight behind the work of Richard Layard and others at CLE on unemployment persistence and tackling the problems of long-term unemployment. He has one important paper on this subject (Pissarides 1992). But, unlike some of his LSE colleagues, Pissarides has not published a unified vision of the workings of the labour market based on his empirical findings, perhaps because many of the results, and the policies which follow from them, may be specific to time or place, rather than having any more universal application.

In addition to all this work on labour markets, Pissarides has a variety of other papers on various other major topics in macroeconomics, including political economy, education policy (e.g. Pissarides 1982) and, more recently, growth with structural change (e.g. Ngai and Pissarides 2007) as well as papers on political economy, consumption, saving and retirement, together with a number of more general articles on macroeconomics.

Given the range and quality of Pissarides's work, it is unsurprising that it has been widely recognised outside LSE. He was the winner of the IZA prize in labor economics (jointly with Dale Mortensen) in 2005, and is currently President of the European Economic Association. He is also a Fellow of the British Academy, the Econometric Society and the Society of Labour Economists, and a research fellow of the Centre for Economic Policy Research (CEPR) and of the Institute for the Study of Labor (IZA, Bonn). He has also been a consultant on labour market issues for the World Bank, the European Commission, the Bank of England and the OECD.

Pissarides has therefore achievements across the whole range of professional economics, in teaching and research spanning pure theory, empirical work and policy analysis. It may be that the Nobel Committee, perhaps mindful of Keynes's assertion that it is ideas which rule the world, have focused their citation on the innovative developments in the theory of markets with search frictions. But the economics profession should recognize the many important contributions over a much wider area.

See Also

- ► Search Theory
- ► Unemployment
- ► Labour Markets

Bibliography

Albrecht, J. 2011. Search theory: The 2010 Nobel Memorial Prize in Economic Sciences. *The Scandinavian Journal of Economics* 113: 237–259.

Alchian, A. 1970. Information costs, pricing and resource unemployment. In *Microeconomic foundations of employment and inflation theory*, ed. E.S. Phelps. New York: Norton.

Beveridge, W.H. 1944. *Full employment in a free society.* London: George Allen & Unwin.

Jackman, R., R. Layard, and C.A. Pissarides. 1983. On vacancies. *CLE Discussion Paper 165*.

Lipsey, R.G. 1960. The relationship between unemployment and the rate of change of money wages in the United Kingdom, 1862–1957. *Economica* 27: 62–70.

Mortensen, D.T., and C.A. Pissarides. 1994. Job creation and job destruction in the theory of unemployment. *Review of Economic Studies* 61: 397–415.

Mortensen, D.T., and C.A. Pissarides. 1999. Unemployment responses to 'skill-biased' technology shocks: The role of labour market policy. *Economic Journal* 109: 242–265.

Ngai, L.R., and C.A. Pissarides. 2007. Structural change in a multi-sector model of growth. *American Economic Review* 97: 429–443.

Phelps, E.S., et al. 1970. *Microeconomic foundations of employment and inflation theory*. New York: Norton.

Pissarides, C.A. 1982. From school to university: The demand for post-compulsory education in Britain. *Economic Journal* 92: 654–667.

Pissarides, C.A. 1984. Search intensity, job advertising and efficiency. *Journal of Labor Economics* 2: 128–143.

Pissarides, C.A. 1985a. Short-run equilibrium dynamics of unemployment, vacancies and real wages. *American Economic Review* 75: 676–690.

Pissarides, C.A. 1985b. Taxes, subsidies and equilibrium unemployment. *Review of Economic Studies* 52: 121–134.

Pissarides, C.A. 1990. *Equilibrium unemployment theory*. London: Blackwell (Second edition. 2000. Cambridge, MA: MIT Press.).

Pissarides, C.A. 1992. Loss of skill during unemployment and the persistence of employment shocks. *Quarterly Journal of Economics* 107: 1371–1391.

Pissarides, C.A. 1994. Search unemployment with on-the-job search. *Review of Economic Studies* 61: 457–475.

Place, Francis (1771–1854)

R. K. Webb

English Radical, born in London on 3 November 1771, died in London on 1 January 1854. Apprenticed to the leather-breeches trade, he developed his radical activism while a member of the London Corresponding Society from 1794 to 1798. A comfortable fortune made as a master tailor after 1799 made possible his second career in politics, beginning with the startling radical victory he engineered in the Westminster election of 1807. Deeply involved in the parliamentary reform agitation of 1830–32, he devised the famous placard, ‘To Stop the Duke, Go for Gold’, intended to prevent the Tories from taking office by forcing a run of the Bank of England. He drafted the People’s Charter in 1838, though he took no part in the later, more extreme phase of Chartism, and was active in the early stages of the Anti Corn Law campaign.

For many years, the library behind his shop in Charing Cross was a gathering place for metropolitan radicals. Introduced by James Mill to Jeremy Bentham around 1809, Place was a vital mediating influence between working-class leaders and radical intellectuals. His sole contribution to economic literature is *Illustrations and Proofs of the Principle of Population* (1822), provoked by William Godwin’s second reply (1820) to Malthus. More sanguine than Malthus about the reform of institutions, he rejected Godwin’s inconsistency and naiveté, His defence of Malthusian principles and methods testifies to his own faith in individual effort and to the improvements in civilization he witnessed in his lifetime. Place launched the first ‘neo-Malthusian’ campaign for contraception, and in 1824–5 was the organizing genius of the successful effort to legalize trade unions and to repeal the ban on export of machinery and emigration of artisans.

See Also

- ► Bentham, Jeremy (1748–1832)
- ► Mill, James (1773–1836)
- ► Utilitarianism

Selected Works

1822. *Illustrations and proofs of the principles of population,* ed. Norman E. Himes. London: Allen & Unwin, 1930. *The autobiography of Francis Place, 1771–1854,* ed. Mary Thale. Cambridge: Cambridge University Press, 1972.

References

Wallas, G. 1898. *The life of Francis Place, 1771–1854*, 2nd ed. London: Allen & Unwin, 1918.

Planned Economy

Alec Nove

'Planning; planned: Intended, in accordance with, or achieved by, a careful plan made beforehand'. This is the Chambers Dictionary definition. Of course in this sense we all plan, whenever we think carefully of what we might do in the future. All economic decision-making relates to the future, since all transactions take time, and in the course of time some circumstances might have changed, and so plans are frequently unfulfilled, or have results different from the original intention.

However, we will have in mind here the deliberate actions of *public* authorities, primarily the state, while referring from time to time also to plans made in the private sector. Plans can be of many kinds. The Soviet version is '*directive* planning' or command planning. The authorities issue binding instructions to subordinate management, telling it what goods and services to provide, from whom to obtain the required inputs, and, as we shall see, much else besides. Then there is *indicative* planning, when the state uses influence, subsidies, grants, taxes, but does not compel. There is also *sectoral* planning, which concerns, for instance, a road network, urban rapid-transit, the coal industry, the national health service. This need not be related to any overall plan for the economy as a whole.

Then there are differences in purpose, reason, objectives. One is to impose the centre's priorities, to replace or combat spontaneous market forces, i.e. deliberately to achieve what would not otherwise occur. This applies most evidently to a war economy, but also to Stalin's economic strategy of the Thirties, with its mass mobilization of material and human resources to create a heavy-industrial base in the shortest possible time. On a less drastic scale these considerations also apply to programmes of rapid development in some Third-World countries, that is, to conscious attempts to transform a country's political economy. In such cases the market is seen as an enemy, to be limited or combated (as in Preobrazhensky's phrase about the battle between 'primitive socialist accumulation and the law of value'), and the same was at least partly true in war economies in the West: prices were fixed, materials allocated, free-market deals in controlled commodities were treated as black-market criminal offences.

However, other kinds of public-sector planning have, or need have, no such hostility to the market, can and do coexist with it. The motive to plan them relates partly to what may be called public goods (e.g. the road network, street lighting, rubbish collection), and partly to externality generating sectors, where the profit-and-loss account of the enterprises concerned constitutes a misleading criterion even on narrowly economic grounds, and/or where private and the more general interests conflict. Examples are many: thus urban public transport, docks, airports, are in the public sector even in the United States. Environmental protection is another important factor: thus in a number of countries deforestation threatens ecological disaster, while in the North Sea it is essential to act to preserve fish stocks, while short-term private profit dictates the cutting of trees and overfishing respectively. There are also natural monopolies, where competition is unnecessary or wasteful: electricity, water, posts, until recently also telephones, are examples; the choice here lies between a regulated private monopoly and state ownership and control. The choice may be influenced also by considerations of public policy. Thus if it is desired to provide a comprehensive postal or telephone service, to supply all houses with pure water, and even remote Scottish islands with electricity, then clearly the public-service aspects must be given some priority: it has always been evident that some of the above activities cannot be profitable.

Some confusion is engendered by the inability to distinguish between *responsibility* for provision of a good or service and the way in which it is provided. Thus, to cite some examples, the public authorities must ensure that city rubbish is collected and disposed of, but this no more requires the rubbish collectors to be public employees than responsibility for road-building requires those who build the roads to be civil servants!

Then there are sectors to which economic profitability considerations may be held not to apply at all: education, health, pensions, are widely held to be the proper subject of planning and provision by public authorities.

Finally, there is the species of planning designed to facilitate and encourage the operation of market-orientated private enterprise. This ranges from infrastructural investment to what is usually called indicative planning, which is not compulsory or imposed, but which helps to fill a most evident gap in the pure free-market doctrine, which is concerned with large-scale investment. Long ago G.B. Richardson (1960) pointed out that, on the assumptions of perfect competition and perfect markets, it is hard to imagine how or why investment should take place, since the profitable opportunity is, by definition, equally visible to all the competitors. Therefore imperfect knowledge and/or collusion, neither of which are in the model, are preconditions for investment. The important role of the state in the success of the South Korean and Japanese export-orientated strategies is inexcusably ignored by the *laissez-faire* ideologists, who can see the success and attribute it wholly to free-market entrepreneurship. Planning of this sort, reinforced by unofficial pressures and fiscal incentives, could be described as a form of stateorganized collusion. In addition there is the role of the state in ensuring macro-balance, or taking counter-cyclical action, which used to be accepted quasi-universally as necessary, though this is now vigorously questioned by the revived *laissez-faire* school, which considers that the economy is basically self-righting.

So only in one of its versions is planning to be seen as in inherent contradiction with the market; in all the others they supplement each other, or plans are actually made operational *through* the market.

Socialist Planning

Socialist planning has a long history. Generations of socialist thinkers, including Marx and his followers, contrasted the deliberate planning that would occur under socialism with the 'anarchy' of capitalism, in which production was for profit, not for use. The 'associated producers' would join together to discuss what is needed and how best it could be provided. As Engels put it, they would compare the useful effect of products with the time necessary to produce them.

Some, for example Kautsky and Lenin, saw a socialist society of the future as if it were one giant enterprise, a single all-embracing factory or office. There would be no 'commodity production', that is, production will be for use, not for exchange. Labour would, when applied, be 'directly social', that is, its use will be validated not *ex post*, through the market, but *ex ante*, by the all-embracing plan, which will express society's needs. Costs would be measured in terms of what was seen as the one ultimately scarce resource, human effort.

Critics, such as Barone and L. von Mises, pointed out some major weaknesses in this approach to socialist planning: the number of calculations required would be enormous, the economic criteria for decision-making would be lacking without meaningful prices. Yet, with but few exceptions, socialists in the marxist tradition persisted in their belief that such planning would be 'simple and transparent' (Marx), that) 'everything would be simple without the so-called value' (Engels); 'capitalism had so simplified the task of accounting and control . . . that any literate person can do it' (Lenin), 'The society of the future will do what is called for by simple statistical data' (Bukharin).

Planning in practice proved to be very complicated indeed. It must be emphasized that it did serve its purpose when that purpose was analogous to that of a war economy: to concentrate resources for the priority objectives determined by the central political authority. When the war did break out, the USSR's survival, after initial military disasters, was in no small degree made possible by the ruthlessly-imposed priority of military requirements. In Western countries too, though in lesser degree, central controls were tight, resources were allocated, and the resultant bureaucratic deformations had much in common with Soviet-type planning. Yet these must be seen as a cost, in the circumstances a necessary cost, of imposing the priorities of war. It was Lange who once likened the Soviet planning system to a war economy, *sui generis*.

In normal times, the priorities become more diffuse, also more numerous. The growth of the economy itself presents new problems and challenges. A Soviet scholar remarked that, if the size of the economy grows six-fold, the number of links to be planned grows to the square of that (or any other) number, i.e. 36-fold, and indeed this can be seen as one expands the number of items included in an input–output table.

The Soviet economy today contains several hundreds of thousands of enterprises, in mining, manufacturing, agriculture, construction, transport, distribution, catering, services. The large number is not due to their excessively small size. On the contrary, it has been argued that Soviet agricultural and industrial establishments are too large, certainly much larger than is the case on average in Western capitalism. Because neither production nor the supply of inputs is based on horizontal, market-type relations, each of these hundreds of thousands of enterprises needs to receive, from some unit in the planning hierarchy, specific instructions as to what to produce, what materials to obtain and from whom, while other plan targets relate to labour productivity, wages, costs, material utilization, investment, technical progress, fuel economy and much else besides. The number of identifiably different products and services, fully disaggregated, has been estimated as upwards of twelve million. The sheer scale of the task of the planners is probably *the* most important source of inefficiency and imbalance. Though Soviet experience shows that a planned economy of this type can function, this same experience strongly supports Barone's conclusion, arrived at in 1908, before there was any practical example to study: it would be difficult but not quite impossible to arrive at a) 'technically' balanced plan, that is, one where the needed inputs match the intended output, but *quite* impossible to see how one could approach an *economic* optimum. Thus it is indeed very hard for those institutions responsible for material allocation to ensure that the needed inputs are provided, but they seldom have the practical possibility or the information to ensure that the inputs are those which are most economical.

This is but one of the difficulties attributable to the sheer scale of the required coordination between multi-million plan-instructions. Academician Fedorenko quipped that next year's plan, if fully checked and balanced, might be ready in approximately 30,000 years time.

It is necessary to distinguish between *long-term* and *current* planning. The long- (or medium) term plan looks forward to the end of a quinquennium, or in some instances as much as fifteen years; thus in 1985 some targets were published relating to the year 2000. These plans are necessarily highly aggregated, and contain broad objectives relating primarily to productive capacity (and so to investment), rather than to the product mix, which will be adapted to requirements which cannot be foreseen in advance in detail. A long-term plan must be balanced in an input–output sense, and planners proceed by so-called material balances for major products, ensuring that planned availability matches planned utilization. These plans are not yet operational, that is, they have no 'addressee': no specific enterprise is instructed to act. Or rather the addressee is the planning and administrative mechanism itself. It is true that there have been proposals, and even decisions, about the need to incorporate enterprises' own quinquennial plans into this process, and indeed to make these plans stable and to relate various norms and incentives to them. However, this has not been possible in practice. Indeed, stable 'micro' plans for five years ahead are surely an impossibility, when even annual plans are notoriously unstable, being altered repeatedly during the period of their currency to cope with the unexpected or to correct errors belatedly identified.

The drafting of the relatively aggregated 'unaddressed' longer-term plans does not present an impossible task, there being only several hundred items. It is the operational annual plan, broken down by quarters and by months, which presents formidable problems. It is drafted in the last few months of the previous year. According to one Soviet source, output plans are made for about 48,000 products, which implies that on average each will contain about 250–300 subproducts or varieties. To go into greater detail would cause inordinate delay. But since each of the 48,000 requires numerous inputs, which must be provided through the allocation mechanism, and since every enterprise must receive specific plan-instructions relating to output and inputs, even in relatively aggregated form the burden on the planners is huge. The essential task of coordination is rendered the more complicated by the fact that responsibility is necessarily shared by numerous separate planning departments and economic ministries.

Centralized Planning

The centralized planning model is based upon the supposition that 'society' (i.e. in practice the planning agencies, under the authority of the political leadership) knows or can discover what is needed, and can issue orders incorporating these needs, while allocating the required means of production so that the needs are economically met. It is worth noting that in some sectors this supposition is close to reality. Thus electricity is a homogeneous product, power stations are interlinked into a grid, information on present and estimated future needs is best assessed at the centre, as it is also in many Western countries. The centre is also the obvious place for decision-making on armaments production. However, a very wide range of goods and services, both producers' goods and consumers' goods, are supplied in a wide variety of types, models, sizes, to serve specific needs. Choice of technique, decisions on new products, possible alternative uses of agricultural land, are matters on which the centre has little relevant information which could serve as a basis for micro-commands. Also it is an evident fact that management possesses vital information as to the production potential of the enterprise, and the planners must rely on an upward flow of proposals and suggestions if they are to issue the correct orders. 'Many if not most commands in a command economy are written by those who receive them', remarked a wise Hungarian, in conversation.

Devolution of authority is thus not only necessary, but inevitably occurs, since plans are frequently late, contradictory, aggregated, and their implementation requires much managerial ingenuity, which frequently has to stretch the boundaries of legality. But the system lacks any criterion for managerial decision-making other than the plan-targets to which management's bonuses and promotion prospects are related. Since prices do not, in either theory or practice, reflect supply-and-demand relationships, relative scarcities or demand intensity, profitability cannot serve as a rational criterion for micro decision-making. Furthermore, because of lack of time and imperfect knowledge, the planners are compelled to proceed on the basis of past performance, introducing the so-called ratchet effect: output targets in the next plan period will be a little higher, costs a little lower, than in the previous period, and indeed all concerned proceed on the assumption that no major changes in past supply or delivery arrangements are likely to occur. It is this which enables the system to function, but Soviet sources understandably criticize these methods, since they are not only conservative, but stimulate undesirable behaviour by management. The latter, judged by plan-fulfilment, seeks a plan easy to fulfil and avoids doing too well in case the following year's target is set too high. Fears of supplies not arriving, and of arbitrary plan changes, also stimulate hoarding of labour and materials, and over-application for inputs.

Attempts to fulfil aggregate plan targets, in roubles, tons, square metres or whatever, engender some familiar distortions, when management produces not for the customer but for plan-fulfilment statistics. This can generate the sort of waste which is typified by the building industry (whose plan is in roubles spent) trying to use the dearest possible materials, and metal goods which are unnecessarily heavy to 'clock up' the necessary plan tonnage. It proves to be remarkably difficult to express a plan for heterogenous products in any unit of measure which does not result in unintended distortions. The weak position of the customer is due to two causes: the supplier is a *de facto* monopolist, and there is a chronic tendency for shortages to occur, which finds expression in a) 'take-it-or-leave-it' attitude on the part of the supplier. But perhaps the most fundamental cause is the one already mentioned: the model requires the centre's plans to incorporate requirements in a degree of detail which is impossible in the complex multiproduct real world, and yet it is these necessarily aggregated plan-targets which serve as the basis for micro-economic activity of enterprises, since they are judged by their fulfilment of these targets.

Initiative is likewise (unintentionally) frustrated. It is not only that management is riskaverse, since risk-taking is not as such rewarded. It is that any new action requires not only motivation but also information and means. Thus innovation, whether in product design or in production methods, is frequently rendered impossible because the required machines or materials are not obtainable, these being allocated by remote bureaucratic offices.

While enterprises are supposed to operate on so-called) 'economic accounting', in fact money and prices generally play a passive role, priority being given to plan-fulfilment indicators. The absence of any built-in incentive to economize has meant the proliferation of compulsory cost-cutting and material-economy plans, which can conflict with the objective of providing what the customer requires. While citizens are free to spend their wages on goods in state shops at state-fixed prices, there is no direct economic link or feedback from these prices to the wholesale prices received by the producing enterprises.

Such a planning system as this becomes increasingly unable to cope with the challenges of what has come to be known as) 'intensive growth', that is, growth based on the more efficient use of scarce resources. However, this same system does give to the political authorities, that is, party and state officials, a high degree of control over material and human resources. There also has developed a kind of informal social contract with the masses: security of employment, toleration of slackness at work, prices of necessities and rents kept low. Any major changes, towards some species of 'market socialism', would thus encounter considerable resistance at all levels of society.

Market Socialism

The Hungarian New Economic Mechanism (NEM), introduced in 1968, sought to overcome the deficiencies of the Soviet model by the limited use of the market mechanism as the basis of current enterprise operations. That is to say, enterprises made their own output plans, based upon negotiated contracts with customers, and purchased their inputs without having to apply for an administered allocation. The 'addressed' current obligatory plan was eliminated. State plans were now to be concerned mainly with investment, that is, with the creation of new capacity and structural change. Prices, market forces, profitability, were to play a major role in guiding the actions of management. However, state-owned enterprises remained under the ultimate authority of economic ministries, and, as also in the Soviet Union, party officials can issue orders on almost any subject.

Hungarian experience can only be seen as a partial test of the viability and effectiveness of the 'market-socialist' model, and this for a number of reasons. One of these has little to do with the model itself: Hungary was hard hit by adverse terms of trade in the Seventies, and the resultant strains led to adverse effects on living standards and to the imposition of tighter controls than was envisaged within the logic of the model, and this included controls over prices. Another 'external' factor was that Hungary trades mainly with other communist-ruled countries, and this trade is predominantly based on annual inter-governmental bilateral deals, a procedure inconsistent with the 'market' logic of the NEM. But there were other problems, which may highlight some contradictions inherent in 'marrying' the principles of market and of socialism. Thus the market requires competition, but there is little competition in Hungarian industry, partly because it is a small country with few producers, but also because of mergers. Competition in turn generates winners and losers, but the commitment to full employment and the pressures from the unsuccessful result in there being no bankruptcies: the loss-makers receive a subsidy, while extra taxes are levied on those judged to be too successful. For all these reasons, the micro-economic logic of the NEM's 'mix' of plan and market has had only limited success.

The success is particularly visible in two sectors: agriculture and distribution (trade, catering, services). In agriculture cooperative (collective) farms are freed from compulsory delivery quotas, freed also from the need to apply to the planners for authority to purchase their inputs (usually they are able to buy them without any permits). There is much more autonomy, much less outside interference, than in the USSR, and also greater flexibility in providing incentives for peasants, and in allowing scope for peasants' private activities as well as for non-agricultural activities of the farms themselves. Since agriculture is notoriously unsuitable as an object of central planning, this is indeed a sector which benefits from reliance on decentralization and the market. Trade and catering benefited both from realistic pricing (persistent shortages of many goods in the USSR were at least in part due to the tendency to underprice them), and also from the legalization of a sizeable private sector: thus many shops and restaurants in Hungarian cities are either privately owned or leased from the state by private operators. Competition has a visible effect on quality and service. Private ('second economy') activities are legal also in construction, repairs, transport (private taxis are allowed) and a range of small-scale manufacturing. In the USSR most of these activities would be illegal, but a sizeable underground second economy exists there also. Thus in Hungary one can observe both the advantages and difficulties which arise when plan and market are allowed to coexist – though of course the particular) 'mix' that exists in Hungary is not the only possible one.

It is noteworthy that Poland and China have formally adopted a model which resembles the Hungarian NEM, though one difference concerns agriculture: in Poland the bulk of the peasantry have remained private smallholders, while in China the 'household responsibility system' introduced after 1979 has effectively de-collectivized the peasantry. In the Polish case the serious economic difficulties which persist have been an obstacle to the implementation of these reforms. In China the resolution adopted in October 1984 explicitly asserts the need to recognize the role of market forces as well as of state planning, and, along with greater freedom for peasant agriculture, petty private trade and ownership have been legalized. This is a 'mix' reminiscent of NEP in the Soviet Union in the early Twenties. However, it is too soon to conclude that the Chinese have a new and durable plan model. One of their leaders remarked that, while managers must be allowed to show enterprise and spread their wings, and they had been confined to too small a cage, insisted that there must be a cage: 'otherwise the bird will fly away'. There appears to be considerable differences of opinion among Chinese party leaders as to the meaning of present policies. Is the use of the market, and the opening to foreign capital, a temporary phase, as NEP was in Russia, with some sort of real socialist planning to follow? Or is the mix between plan and market a long-term model of socialist planning? The rapid growth of income inequalities, the corruption of many officials, a speed-up in inflation, could lead to a counterattack, to the reimposition of more central planning. This is not the place to speculate on such matters, only to note that the Chinese are still seeking their own model.

Yugoslavia's combination of plan and self-management was also based in principle on the use of the market mechanism. The micro-economy was to function on the basis of contractual relations between self-managed enterprises, guided by material advantage and by realistic free-market type prices. The problems related to self-management are treated elsewhere (see ▶ Market Socialism). Yugoslavia's economy has run into serious difficulties, not least because the necessary minimum degree of central planning was absent. Tinbergen wisely remarked that under conditions of self-management, 'it can be convincingly shown that in an optimum order some tasks must be performed in a centralized way and cannot therefore be left to the lowest levels (Tinbergen 1975).

Part of the problem was republic–regional fragmentation, complicated by a long history of local nationalisms, so that each republic tended to make its own investments, to keep its own earnings from exports, to run its own finances, which helped to disintegrate the economy of what is, after all, a small country. There is a moral here of wider application about regional planning powers; a regional authority will tend to divert resources for the use of its own region, even if it harms others, if it has the power to do so. But this is but one aspect of a more general problem: the interests of the parts do not necessarily add up to the interests of the whole. There are economies (and diseconomies) of scale, and externalities, which cannot be ignored. Furthermore the self-management model itself tends to encourage excessive income distribution and discourage labour-intensive investments, a situation which can and did give rise to serious unemployment combined with accelerating inflation. The latter was also due to lack of adequate control over credits issued by the (numerous) banks, and to what for several years was a negative real rate of interest.

Yugoslav experience does not prove that either self-management or the market mechanism were wrong. It does strongly point to the need of economic powers at the centre, not only to ensure macro-economic balance but also to devise and enforce the ‘rules of the game’ for the micro-economy. It also demonstrates the limitations and dangers of ‘socialist *laissez-faire*’. If the USSR’s economy is stifled by allembracing central controls, then Yugoslavia shows the consequences of having no systematic central controls at all.

This criticism can be extended to some early models of a decentralized socialist economy, such as that of Oskar Lange and Abba Lerner. These do contain a Central Planning Board, but it is imagined as functioning only via the fixing of parametric prices, to which management is supposed to react in accordance with the best neoclassical principles. Intended to show, in reply to critics (notably Mises), that socialist planners do not require to solve innumerable simultaneous equations, Lange’s counter-model contained neither growth nor indeed any plan at all. Nor, or course, did the world of Mises. What was shown was that equilibrium with efficient allocation would be possible, on the abstract assumptions common to both protagonists. It is worth reminding oneself of Kornai’s dictum: few indeed are those who take decisions on the basis only of information about price (especially when, in taking investment decisions, the relevant prices are those of the future).

Those critics of socialist planning who emphasized the alleged impossibility of solving too many simultaneous equations had grounds for alarm when the computer, programming, input–output techniques, appeared to make the impossible possible. After all, whereas in a capitalist planless society there was and could be no operational objective function, a centrally planned economy could – it might be supposed – use the new techniques to arrive at the most economically efficient way of achieving the objectives defined by the supreme political authority, which is simultaneously in command of the economy. Indeed some members of the Soviet mathematical school explored in very interesting ways how this might be attempted, and Kantorovich, who received the Nobel prize for his pioneering work on linear programming proposed a system of plan-valuations which could be used in calculations designed to achieve optimal allocation (and had to defend these valuations from criticism from dogmatic defenders of the labour theory of value).

It turned out that progress along this route was disappointingly slow. We can now see more clearly why. Firstly, the ‘objective function’ proved to be operationally indefinable, despite efforts by able mathematical economists to define it. What could be the objective basis for an optimal plan, what could be the criterion by which to judge if any given plan were optimal? The objectives of the political leadership cannot serve as such a criterion, since (as one Soviet economist remarked) it seeks advice as to what the plan objectives should be, and would not thank those economists who replied that its wishes were their criterion. Any real society generates numerous inconsistent objectives, and in a one-party state these are also present, and find expression within the one party. Then the ‘curse of scale’ must again be emphasized. Botvinnik, the former world chess champion, estimated that the number of possible moves in a chess game exceed substantially the number of words spoken by all human beings since the Pyramids were built. A chess board has only 64 squares, and rules of the game are known. An economy or a society has many more, and the human ‘pieces’ play different games and dispute about the rules. So even if one day a chess grandmaster might have trouble beating a chessplaying computer, the idea that computers could replace markets and make Soviet-type centralized planning ‘efficient’ is surely a chimera. It is true that computers can aid the centre in making calculations. They have numerous uses at micro, i.e. decentralized levels, as a source of data, or in design bureaux, etc. However, one can scarcely imagine that the centre can administer through a computerized programme a fully disaggregated micro-plan for millions of products, distributed among hundreds of thousands of enterprises. Not only would there be too much information to handle (and check), but decisions involving quality, or judgement as to uncertain outcomes can hardly be left for computers. Scale is also a hindrance to the use in practice of prices based on central computerized programmes (the ‘objectively determined valuations’ of Kantorovich). At operational disaggregated level there is no such thing as *the* price of ‘agricultural machinery’ or ‘ball-bearings’, or ‘footwear’: there are hundreds or thousands of different products under each of these heads, which need to be provided, and priced, for different requirements or preferences.

Plan and Market

Plan and market have been seen as incompatible opposites, both by dogmatic socialists and by dogmatic anti-socialists. However, a strong case can be made for the proposition that a mix of the two is essential in any modern society. True enough, a long list can be made of distortions and deficiencies directly attributable to planning. Disastrous indeed have been some comprehensive redevelopment schemes devised by well-meaning urban planners, and some of the housing has later had to be dynamited. Planning foreign trade in a number of countries, especially in the Third World, has been a means of personal enrichment for those entitled to issue import licences. Development plans have sometimes been grandiose and wasteful. From these and similar experiences some have drawn the conclusion that planning is ‘bad’, that reliance on the market mechanism will provide the right answers to all economic problems, and that state intervention should confine itself to controlling the money supply and to providing a minimum range of so-called public goods, such as defence and lighthouses. Conversely, socialists see that the operation of the free market generates excessive income inequalities, gives rise to monopolistic abuses, to trade-cycles, to unemployment. The market inspires acquisitiveness, substitutes conflict (between classes, and also between competitors) for the desired harmonious cooperation.

Yet both sets of dogmatists appear to be mistaken. The evils which they have noted do indeed exist, and require to be explicitly recognized and combated. The difficulties faced by centralized marketless socialism have already been discussed at length, and it is hard to see how decentralization could be envisaged without some sort of market mechanism which would link the parts together. *Laissez-faire*, the belief that virtually all public-sector planning or provision is either harmful or unnecessary, ignores much of what did or does happen in the real non-textbook world.

Investment is clearly one relevant sector. Given the degree of uncertainty facing private investors, their understandable desire for security, the attraction of high interest rates (and the negative effect of such high rates on would-be borrowers), it would seem to be a remarkable act of faith to imagine investments, especially in the longer term, would be rational, let alone optimal. Various forms of indicative planning, reinforced by the state’s own investment plans (e.g. in infrastructure), become an important contribution to guiding private investment decisions. As already mentioned, the South Korean government played a key role in the process of developing highly successful exporting sectors. If interest rates are (say) 15 per cent, whose private concern should it be to think about (for instance) the consequences for Great Britain of the exhaustion of North Sea oil or gas supplies by the end of the century? It requires ideological obstinacy of a high order not to see that an energy plan might be desirable, in the national interest. The devotees of 'methodological individualism' go so far as to assert that there *is* no national interest, distinct from the individual interests of the citizens. Even on so extreme an assumption it must still be recognized that individual or sectoral interests can conflict with one another; the elementary example of many people wishing to park cars in a narrow street is but one of many instances when people literally get in each others' way, and public authority has to sort out the mess. One returns, too, to examples cited earlier concerning external effects. Docks, airports, rapid-transit systems, have wide-spreading effects – on industrial profitability, property values, congestion, etc. – which do not show up in their respective profit-and-loss accounts. It does seem absurd to assert that the Washington or Montreal (or Moscow, or Munich, or Budapest) metros should not be part of a transport plan for their respective cities, or should not be provided because – as is the fact – they do not 'pay'. But the) 'methodological individualists' are plainly mistaken. In virtually any institution, from the state to a firm or a university, it is frequently possible for the perceived interest of the part to conflict with that of the whole. While it is too complex and time-consuming to attempt to 'internalize' all externalities, it is essential to try to identify contradictions and conflicts of interest when these are important, and not to evade the issue by pretending that – with appropriate legal and institutional arrangements – they will not exist. State intervention is one form of dealing with these problems, in the general interest.

Businessmen, especially at times characterized by uncertainty and high interest rates, have a short-time horizon. Thus Nobel laureate Wassili Leontief wrote: 'Our [US] business man investor expects to get back his capital in about four-and-a-half years. So really he is not very worried about what will happen beyond these four and a half years' (*The Federalist*, March 1985, p. 66). This is not necessarily in the long-term interest of the firm, let alone of the entire economy or of society.

Some extreme anti-planners need reminding of the fact that trade-cycles existed even when trade-union powers were minimal, that chronic unemployment may be as irrational a waste of resources as anything that happens in a centrally planned economy. The notion that labour markets 'clear' but for remediable imperfections is surely a myth derived from general-equilibrium analysis. Real competition *requires* unused capacity, necessarily involves winners and losers, otherwise how could competition actually proceed? This is apart from the serious danger of technologically induced long-term unemployment, which may pose a major threat to overall stability. Yet to combat unemployment by an expansionary policy can engender accelerating inflation unless consideration is given to an incomes policy, itself part of a plan.

It is true that, in the effort to plan and control, major errors have been and could be committed. However, to take one last example, the fact that dreadful mistakes in town-planning have occurred does not prove that no town-planning powers should reside in public authorities.

The whole subject remains highly controversial, and ideologies of both left and right heavily influence both policies and theoretical formulations. At present in many Western countries it is the advocates of planning that are fighting a rear-guard battle.

See Also

- ▶ Command Economy
- ▶ Indicative Planning
- ▶ Market Socialism

Bibliography

Nove, A. 1983. *The economics of feasible socialism*. London: Allen & Unwin.

Richardson, G.B. 1960. *Information and investment*. Oxford: Oxford University Press.

Tinbergen, J. 1975. Does self-management approach the ultimate order? In *Self-governing socialism*, ed. B. Horvat et al., 226. New York: IASP.

Planning

Rajiv Vohra

Keywords

Convexity; Decentralization; Excess demand; Increasing returns; Informational efficiency; Lange, O. R.; Marginal rate of substitution; Monotonicity; Planning; Shadow pricing; Socialist calculation debate; tâtonnement

JEL Classifications

D6

Formally, planning in an economic context can be identified with a constrained maximization problem. The objective, whether it is simply social welfare or multiple individual utilities, is maximized subject to the resource and technological constraints. It needs to be emphasized that the planning problem is not simply one of characterizing the solution to the maximization problem but also of defining a computational procedure to obtain the solution. A planning process can be defined as an iterative procedure which, through successive approximations, finds a solution to the maximization problem.

The literature on planning processes goes back at least to the debate of the 1920s and 1930s on the possibility of economic calculation in a socialist state. While the formal versions of the welfare theorems, as presented by Arrow (1951) and Debreu (1954), were not available then, it was fairly well recognized that the competitive mechanism would, in equilibrium, satisfy the marginal conditions in terms of the equality of prices and the relevant rates of substitution and that this would constitute an efficient method of allocating resources. In what seems, at least in retrospect, to be an argument one may well be tempted to make if one was aware of the second welfare theorem, Mises (1922) argued that since the markets for capital goods, and hence their prices, would not exist in a socialist economy, it would be impossible for such an economy to allocate its resources rationally. However, Pareto (1897), in comparing the market to a computing machine, had already pointed out that a procedure similar to the competitive process of the market could be used to determine a plan. His argument had been further elaborated by Barone (1908). The focus of Mises's criticism was somewhat changed by Hayek (1935) who did not rule out the theoretical possibility of a planned economy being able to allocate resources rationally. The scepticism was centred around the ability of the planning authority, say the Central planning Board (CPB), to solve the 'hundreds of thousands' of equations necessary to achieve the objective. Partly in response to this criticism, iterative processes were presented, in what are now famous papers by Taylor (1929) and Lange (1936–7), to show that a planned economy could allocate resources in much the same way as the competitive system. They formalized a planning process which would follow the competitive rules to allocate resources; the trial and error method for finding the optimal allocation was similar to Walrasian tâtonnement. The arguments presented by the sceptics were turned on their head; the planned economy could play the competitive game just as well as the market, perhaps better.

While, in the classical environment, a process which imitates the competitive market has the clear advantage of leading, in equilibrium, to a Pareto optimal allocation, the dynamic properties of such a process were analysed much later. Samuelson (1949) showed that in a linear economy, such a mechanism led to indefinite oscillations. Arrow and Hurwicz (1960) rigorously formalized Lange's process, for an economy with a single utility function, and showed that strict concavity of the utility function and the technological constraints were crucial in establishing the convergence of the dynamic process.

In the subsequent development of this literature considerable attention has been paid to developing processes which converge to an optimal plan. Other criteria for comparing different processes have also been formalized (see for example Hurwicz 1960, and Malinvaud 1967) and we shall discuss these in more detail in section "The Formal Model and Definitions". At this stage it is, however, worthwhile to point out that a planning process which mimics the competitive process has considerable appeal. In the classical environment it leads to an allocation which is Pareto optimal. It also retains the attractive informational processing properties of the competitive mechanism; the CPB is not required to collect all the information on the economic environment nor does it need to solve the entire programming problem by itself since various stages of the optimization process are conducted at the individual level. Subsequent literature on planning processes has, quite justifiably, concentrated on processes which are in some sense decentralized. Processes applicable to non-classical environments have also been formulated.

There is also a considerable literature on general allocation processes in which the CPB is not assigned a distinguished role (see for example Arrow and Hurwicz 1977). In this article we shall concentrate only on planning processes. In particular, we consider an economy with many firms and a CPB. Except for the section on public goods where we consider many consumers, the CPB is assumed to have the objective of maximizing a single utility function. Section "The Formal Model and Definitions" will set out the model and the criteria which may be used to compare different processes. Processes designed for the classical environment are considered in section "The Classical Environment". Sections "Increasing Returns" and "Public Goods" deal respectively with economies with increasing returns and with public goods. Due to limits on space, we shall not deal with other non-classical environments that have also been studied in the literature on allocation process (see for example Section III of Hurwicz 1973). Another important aspect of planning which is not covered here is that of incentive compatibility. Moreover, the discussion is not intended to cover all the details of the processes under consideration and the reader may find it useful to consult the cited papers. Notable among the surveys in this area are Heal (1973), Hurwicz (1973) and Tulkens (1978).

The Formal Model and Definitions

We shall consider an economy with k commodities, indexed by l, and $n + 1$ agents, n firms and the CPB. Agents will be indexed by

$$i, i = 1, \ldots, n, n + 1.$$

We shall also find it convenient to index the firms by j, $j = 1, \ldots, n$. Firm j's technology is represented by production set $Y^j \subseteq R^k$. The environment of firm j is simply $e^j = Y^j$. The CPB has a continuous utility function U: $X \rightarrow R$ where X is the consumption set. The aggregate endowment of the economy is denoted $\omega \in R^k$. The economic environment of the CPB is $e^{n+1} = (X, U, \omega)$. The economy can be described in terms of its environment $e = (X, (Y^j), U, \omega)$.

D.1. A program (x, y) consists of a consumption plan $x \in X$ and a collection of production plans $y = (y^j) \in Y = \Pi_j Y^j$.

D.2. A program (x, y) is said to be feasible if $x = \Sigma_j y^j + \omega$.

D.3. A program (x, y) is said to be Pareto optimal if it is feasible and there does not exist another feasible program $(\bar{x}, \bar{y})$ such that $U(\bar{x}) > U(x)$.

A planning process is an iterative process in which messages are exchanged between the firms and the CPB. Agent i chooses a message m^i from a set M, taking into account the environment and the messages received in the previous period. Let m_t^i refer to agent i's message in time period t and

$$m_t = \left(m_t^{\,1}, \ldots, m_t^j, \ldots, m_t^{\,n+1}\right).$$

The response of agent i may then be defined in terms of a response function $f^i : M^{n+1} \rightarrow M$ where M^{n+1} refers to the $n + 1$ fold Cartesian product of M and

$$m^i_{t+1} = f^i(m_t; e).$$

An equilibrium message is simply defined as a stationary message. The equilibrium of the process is determined by an outcome function h which translates the equilibrium message into the equilibrium program or plan. We can now formally define these concepts:

D.4. Given an environment e, a *planning process* is defined as $\pi = (M, f, h)$ where $f : M^{n+1} \rightarrow M$ and $h : M^{n+1} \rightarrow X \times Y$.

D.5. An *equilibrium message* for a process π is an $m \in M^{n+1}$such that $m = f(m; e)$.

D.6. An *equilibrium program* (or an *equilibrium plan*) for a process π is a program (x, y) such that $(x, y) = h\ (m; e)$ and m is an equilibrium message.

We shall now discuss some of the desirable properties that a planning process may have. These properties may be broadly classified in terms of the performance of the process and its informational efficiency. We begin by presenting the performance criteria introduced in Malinvaud (1967).

Clearly convergence to a Pareto optimal allocation is a requirement that any planning process ought to satisfy.

D.7. A process π is said to be *convergent* if an equilibrium program exists and is Pareto optimal, that is, as $t \rightarrow \infty$ $\text{Lim}_t h(m_t; e)$ is Pareto optimal.

Malinvaud also stresses the importance of the following properties which may, in practice, be even more important if the process needs to be terminated before equilibrium is reached.

D.8. A planning process $\pi = (M, f, h)$ is said to be feasible if $f(m, e)$ and $h(m, e)$ are nonempty and $h(m, e)$ is feasible for all $m \in M^{n+1}$.

D.9. A planning process π is said to be *monotonic* if $U(x_{t+1}) \geq U(x_t)$ for all t, where x_t is the consumption plan corresponding to $h(m_t; e)$. It is *strictly monotonic* if it is monotonic and $U(x_{t+1}) = U(x_t)$ implies that $h(m_t; e)$ is Pareto optimal.

Hurwicz (1960, 1969) formalized the notion of informational efficiency associated with a process. His definitions are applicable to general allocation processes in which a CPB is not assigned a distinguished role and we shall suitably modify his concepts to apply specifically to planning processes. The definitions which follow are aimed at formally defining a decentralized process, a definition which is intended to include but not be synonymous with the competitive process. An important characteristic of the competitive system is that initial information is dispersed among the agents: firm j knows only its own environment Y^j while a consumer knows only his or her utility function and endowment. A process in which the ith agent's response functions depends only on e^i is said to be *external*. A process is *anonymous* if the agents do not know the source of their messages. Since there is only one planning authority, this requirement may not be relevant for the firms; if certain kinds of messages are transmitted only by the CPB the firms would know the source of these messages. As far as the CPB is concerned, it would be desirable if messages did not have to be identified with particular firms. In particular, if the aggregate response of the firms is all that the CPB needs to determine its message, this must be considered a significant advantage. Clearly, this would be a stronger requirement than anonymity and a process satisfying this requirement will be called *aggregative*. Another informational requirement that Hurwicz (1969) imposes on a decentralized process is that the message space M be R^k. Calsamiglia (1977) considers a somewhat less restrictive condition on the amount of information that needs to be transmitted. He defines a process to be *point valued* if M is some finite dimensional Euclidean space.

D.10. A process $\pi = (M, f, h)$ is *informationally decentralized* if it is external, aggregative and point valued, that is if

$$m^j_{t+1} = f^j\left(\sum_{)j(} m^j_t, m^j_t, m^{n+1}_t; e^j\right),\ m^{n+1}_{t+1}$$

$$= f^{n+1}\left(\sum_j m^j_t, m^{n+1}_t; e^{n+1}\right)$$

and

$$M = R^s,$$

where $\Sigma_{)j(}m^j$ refers to the summation cross the messages of all except the j-th firm and s is a positive integer.

While most of the planning processes in the literature are external and aggregative, many of them are not point-valued in the above sense. In particular, Malinvaud's (1967) process is one in which the agents transmit point-valued messages in every time period but the response of the CPB depends also on the messages received in the past. Such a process would not be informationally decentralized according to the above definition; however, there is something to be said for making a distinction between messages and memory, and between a process in which messages at each point in time are infinite dimensional and one in which finite dimensional messages are transmitted but the CPB has a memory of past messages. Processes of the later variety have also been termed decentralized and while this may not be unreasonable, it has led to some confusion (see Cremer 1978).

The Classical Environment

In this section we discuss planning processes designed for the classical environment in which there are no externalities, production sets are convex and the utility function is quasi-concave. In this setting, the competitive allocation has the attractive welfare properties that it is Pareto optimal and any Pareto optimal allocation can be sustained as a competitive allocation with a redistribution of initial resources. In an economy with a single utility function, a competitive allocation is unambiguously optimal.

We begin by considering Lange's process as formalized by Arrow and Hurwicz (1960). As mentioned earlier, the Lange–Arrow–Hurwicz (LAH) process is closely related to the Walrasian tâtonnement. Arrow and Hurwicz consider the following process: the CPB announces prices p and the firms choose profit maximizing production plans. The CPB computes a consumption plan to maximize $U(x) - px$. The prices are then varied in proportion to excess demand.

Arrow and Hurwicz formulate the planning problem as a programming problem and apply the gradient method (or method of steepest ascent) to find its solution. They formulate their process in continuous time in the activity analysis framework. We now formally describe the LAH process in its discrete version, as transposed by Malinvaud (1967) to the model presented in section "The Formal Model and Definitions" above. Given prices p, firms choose their profit maximizing plans and the CPB chooses the consumption plan which maximizes $U(x) - px$. The price of a commodity is then increased by an amount proportional to its aggregate excess demand, the coefficient of proportionality being a positive constant ρ provided this change does not make the price negative. The responses of the agents can now be defined formally:

(i) for all $z \in Y^j\}j = 1, \ldots, n, .$
(ii) $x_t = \{x_t \in X|\ U$
$(x_t) - p_{t-1}x_t \geq U(z) - p_t z$ for all z, $\in X\}$,
(iii) $p_{t,l} = \max\Big\{0, p_{t-1,l} + \rho\Big(x_{t-1,l} - \Sigma_j y^j_{t-1,l}$
$-\omega_l)\}\ l = 1, \ldots, k.$

It is clear that in order for the process to be convergent, a Pareto optimal allocation must exist. This in turn will be guaranteed if, for example, all the production sets and the consumption sets are compact. The assumption that the above mappings are all single-valued, that is, there is a unique production plan that maximizes profits for each firm, given *p* and a unique consumption plan that maximizes $U(x) - px$, also turns out to be important for the convergence properties of the process. While the process in continuous time is convergent (see Theorem 12 in Arrow and Hurwicz 1960), Uzawa (1958) showed that the discrete version of the LAH process converges only approximately. The following result is the version presented in Malinvaud (1967).

Theorem 1 If there is a unique Pareto optimal allocation $(\overline{x}, \overline{y})$ and the functions defined by (i) and (ii) are single-valued, the process defined by (i), (ii) and (iii) is approximately convergent in the following sense: for any $\varepsilon > 0$ there exist ρ_0 and t_0 both depending on ε such that if $\rho \leq \rho_0$ then for $t \leq t_0$ the distance between $h(m_{t;e})$ and $(\overline{x}, \overline{y})$ is no greater than the distance between $h(m_{t-1}, e)$ and $(\overline{x}, \overline{y})$ and for $t \geq t_0$ the distance between $h(m_t; e)$ and $(\overline{x}, \overline{y})$ is no greater than ε.

It is easy to see that this process is not feasible since, out of equilibrium, aggregate excess demand for some commodities may be positive. There is also the problem that, since the function $\rho(\varepsilon)$ is not known, it is not possible, given some ε, to choose the value of ρ to be most efficient.

Malinvaud (1967) proposes two other processes which are feasible, monotonic, and convergent but are not decentralized according to D.10 since the CPB is required to remember the messages conveyed by the firms in the past. Malinvaud's first process is designed only for a linear economy and is based on Taylor's (1929) proposal. Each firm is assumed to have a set of fixed coefficient techniques that can be operated under constant returns to scale. The CPB announces prices corresponding to which firms respond with a cost minimizing technique. The CPB then solves the open Leontief model to obtain prices which would make firms' proposed techniques earn zero profits and a consumption plan which maximizes utility at these prices. This process is then shown to satisfy Malinvaud's criteria under certain conditions. Malinvaud's second process covers a more general environment and we shall now discuss this in somewhat more detail.

This process is an application of the Dantzig and Wolfe (1961) decomposition algorithm to the planning problem. The CPB builds up an approximation of the firms' production sets based on messages received from them in the past. At each stage firms reveal their profit maximizing production plans, given the prices conveyed by the CPB. Assuming that all the production sets are convex, the CPB can construct a subset of a firm's production set by taking the convex combination of all the production plans revealed by that firm in the past. The CPB then solves the programming problem of maximizing utility subject to the resource constraints and the technological constraints as given by its construction of the firms' production sets. In the next stage the shadow prices obtained from the programming problem are announced as prices and the process continues. For the process to start it is assumed that the CPB has initial information about at least one feasible production plan for each firm.

We can now define the process formally. Let Δ denote the $k - 1$ dimensional simplex and $Y_t^j = \mathrm{Con}\left(y_t^j, y_{t-1}^j\right)$ where Con denotes convex hull. Let $(\overline{x}_t, \overline{y}_t)$ denote the allocation which solves the programming problem at *t*, that is,

$$\begin{aligned}(\overline{x}_t, \overline{y}_t) = \{(x, y) \in XxY | x_t \\ \leq \sum_j y_t^j + \omega \text{ and } U(x_t) \\ \geq U(z) \ \text{ for all } z \in \sum_j \text{ con } \left(Y_t^j\right) + \omega \}.\end{aligned}$$

We shall say that $p \in \Delta$ supports the allocation $(\overline{x}_t, \overline{y}_t)$ if

(a) $U(x) \geq U(\overline{x}_t)$ implies $px \geq p\overline{x}_t$
(b) $y \in Y^j$ implies that $py \leq p\overline{y}_t^j$ for all *j*.

The process is defined by the following equations:

(i) $y_t^j = \{y \in Y^j | py \geq p\overline{y}$ for all $\overline{y} \in Y^j\}, j = 1, \ldots, \mathrm{n}.$
(ii) $p_t = \{p \in \Delta | p \text{ supports } (\overline{x}_t, \overline{y}_t)\}.$

The plan at stage *t* is simply defined as $\left(\overline{x}_t, \overline{y}^j{}_t\right)$.

The following assumptions are sufficient for this process to satisfy Malinvaud's criteria. (A1) X is closed, convex and bounded from below. $U(x)$ is continuous, quasi-concave and locally non-satiated (A2) Y^j is convex and compact for all j. (A3) the CPB knows a feasible program (x_1, y_1). We can now state:

Theorem 2 (Malinvaud 1967) If (A1), (A2) and (A3) are satisfied the process defined by (i) and (ii) is feasible, monotonic and convergent.

To see that this process is feasible notice that, given (A2), (i) always has a solution and, given (A1) and (A2), the programming problem, for any t, has a solution and this solution constitutes the plan for that time period. We can appeal to the second welfare theorem (see, for example Theorem 6.4 in Debreu 1959) to assert that (ii) also has a solution. Monotonicity is an obvious property of this process since the constraint sets in the programming problem of time t are contained in the constraint sets of time $t + 1$. Convergence is established by considering a limit argument, using the fact that all the plans lie in a compact set. We refer to Malinvaud (1967) for the proof.

As the above theorem shows, this process has better performance properties than the LAH process. However, its information requirements are much stronger. The CPB is required to have memory and to know of a feasible allocation. It also solves a rather complicated programming problem at each stage. Moreover, to implement the plan the firms are instructed to follow the production plans computed by the CPB. While these plans are consistent with profit maximization, a specific instruction has to be issued to each firm. But this problem can be avoided in the simpler case where all production sets are strictly convex. In this case, the equilibrium plan can be implemented simply by announcing shadow prices and letting the firms find their unique profit maximizing production plans.

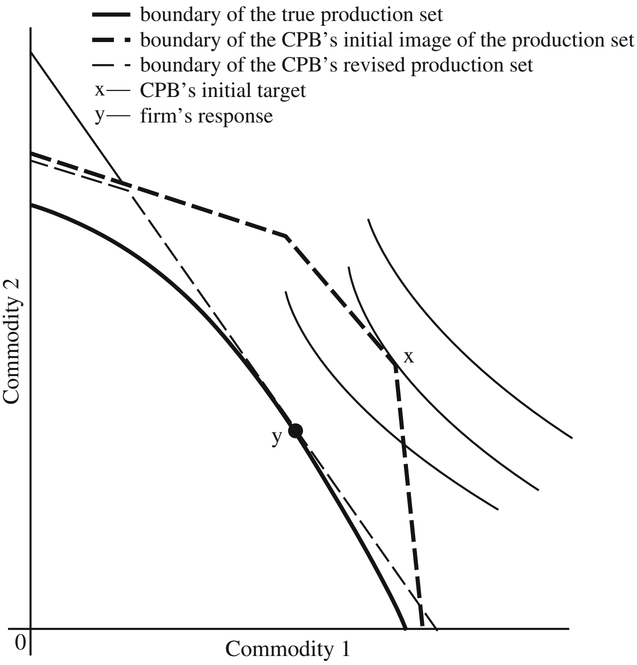
Weitzman (1970) proposed a process which is in a sense a dual of Malinvaud's process. The CPB has a belief about a firm's production set, which is not necessarily correct, and, given these imaginary production sets, it solves the programming problem and provides each firm with a production plan as a target. If the firm finds that this target is not feasible it responds with an efficient plan and a corresponding marginal rate of substitution. The CPB then constructs a new production set which is the intersection of its previous one with the half space determined by the firm's announced efficient plan and marginal rate of substitution. The CPB again solves the programming problem and announces new targets (see Fig. 1). Not only is this process convergent, if the production sets are polyhedral, convergence is achieved in a finite number of steps. However, it is not feasible since the CPB's targets may not be feasible for the firms.

Another process which uses production quotas rather than prices as signals is one due to Kornai and Liptak (1965). Their process is formulated for a linear economy in which the CPB's utility function is separable among the firms' outputs. The CPB allocates resources to the firms which respond with rates of substitution and the CPB reallocates resources in response to the value of the allocated resources at the shadow prices. They model the interaction between the CPB and the firms as a game and show that the process is convergent.

Increasing Returns

Heal (1969) proposed a non-price gradient process which locates local maxima even for economies with increasing returns. The CPB allocates the inputs among firms which then respond with efficient output levels and marginal productivities. The CPB then reallocates the inputs towards the firms with higher marginal productivities. The process can be most easily understood in the simple setting in which all firms produce an identical output using m primary resources. Departing from our notation of section "The formal Model and Definitions", we shall denote by y_i the amount of output produced by firm i and by f_i firm i's production function. The amount of input j used by firm i is denoted x_{ij} and the technological constraints may be stated as follows,

Planning, Fig. 1


boundary of the true production set
boundary of the CPB's initial image of the production set
boundary of the CPB's revised production set
x— CPB's initial target
y— firm's response
Commodity 2
Commodity 1
0
x
y

$$y_j = f_i(x_{i1}, \ldots, x_{im})\ i = 1, \ldots, n,\ \ x_{ij} \geq 0, \text{for all } i \text{ and } j.$$

Let R_j denote the aggregate endowment of the jth resource. The resource constraints can be stated as

$$\sum_i x_{ij} \leq R_j \text{ for all } j.$$

The objective of the planning process is to find $((x_{ij}))$ to maximize $\Sigma_i y_i$ subject to the technological and resource constraints. Let f_{ij} denote firm i's marginal productivity of the j-th input and let a dot over a variable denote its rate of change. The process starts with the CPB allocating x_{ij} to the firms subject to the resource constraints. The CPB then raises the allocation of input j to a firm if its marginal productivity is greater than a certain average productivity and lowers it if it is lower than the average, subject to the non-negativity constraints. Formally, the rate of adjustment is determined by the following equations

$$\dot{x}_{ij} = f_{ij} - \mathrm{Av}\left(K_j\right)f_{ij}, \text{if } i \in K_j\ \ 0\ \ \text{other wise},$$

where $\mathrm{Av}(K_j)f_{ij}$ denotes the average of f_{ij}'s contained in the set K_j. The set K_j is constructed (see Heal 1969) so that the non-negativity constraints are not violated in applying the adjustment equations and it satisfies the following property

$$\mathrm{K}_j = \left\{ i \middle| x_{ij} > 0 \text{ or } x_{ij} = 0 \text{ and } f_{ij} > \mathrm{Av}\left(K_j\right)f_{ij} \right\}.$$

K_j includes firms with positive allocations of input j or firms with a zero allocation but a marginal productivity higher than the average.

We can now state the following theorem, which applies to the simple model we are considering but also extends to the more general case where firms produce different commodities.

Theorem 3 (Heal 1969) If all f_i have continuous, finite first derivatives and the initial allocation is feasible, the process defined above is feasible and monotonic. Moreover, every limit point of the process satisfies the necessary conditions for Pareto optimality. If the initial allocation is not a local minimum, then the limit points are to local minima.

To see that the process is feasible, notice that

$$\sum_i \dot{x}_{ij} = \sum_{i \in K_j} \left[f_{ij} - (1/K_j) \sum_{i \in K_j} f_{ij} \right] = 0, \text{for all } j.$$

Thus, if the initial allocation is feasible so are all other allocations. To establish monotonicity, we consider

$$\dot{y} = \sum_i \sum_j f_{ij} \dot{x}_{ij}$$

which can be written as

$$\begin{aligned} \dot{y} &= \sum_j \sum_{i \in K_j} f_{ij} \left[f_{ij} - (1/|K_j|) \sum_{i \in K_j} f_{ij} \right] \\ &= \sum_j \sum_{i \in K_j} \left[f_{ij} - (1/|K_j|) \sum_{i \in K_j} f_{ij} \right]^2 \geq 0. \end{aligned}$$

Thus, $\dot{y} \geq 0$ and $\dot{y} = 0$ if and only if $f_{ij} = f_{kj}$ for all i and $k \in K_\mathrm{j}$ for all j. It is easy to see that the equality of f_{ij} for all i in K_j is the necessary condition for optimality. This implies that y increases monotonically except when the necessary conditions for optimality are satisfied. In particular, if the initial allocation is not a local minimum, the equilibrium allocation, arrived at through monotonic increases, cannot be a local minimum. Hori (1975) showed that the convergence to a point of inflection is unlikely in a well defined sense. A discrete version of this gradient process would also be approximately convergent in the sense described in Theorem 2 above.

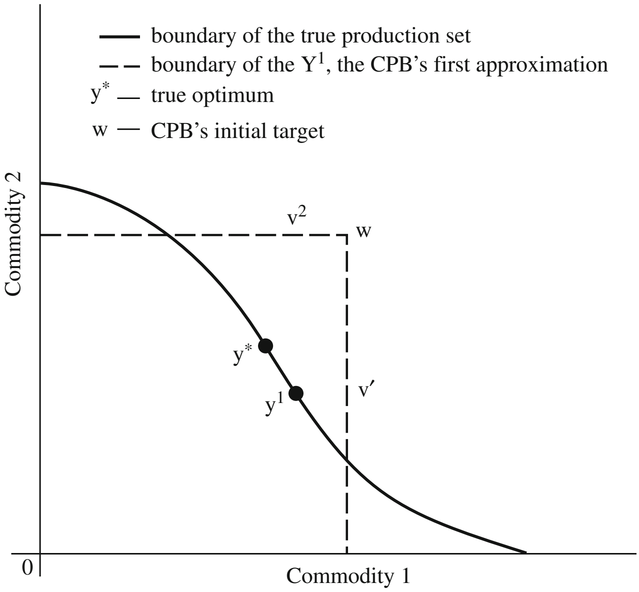
Since this process requires the CPB to respond with allocations to firms, the informational requirements are much stronger than those of a price guided process in which a common price vector is given out to the entire production sector. In the general case where firms produce many commodities the CPB uses marginal valuations not only to allocate inputs but also output combinations to each firm (see Heal 1973, ch. 8). However, unlike the Malinvaud or the Weitzman process, the CPB is not required to have a memory. It is also possible to modify this process to take advantage of the informational efficiency which is characteristic of the price guided processes. Such a mixed planning process was formulated by Heal (1971) and is similar to one proposed by Marglin (1969). In Heal's (1971) process the CPB allocates resources to the firms and also provides them with prices of the final goods. The firms inform the CPB of their profit maximizing output bundles and also of the marginal productivity of the inputs. The CPB reallocates inputs as in the previous process and announces new output prices which reflect the marginal rates of substitution in consumption. The performance of this process is similar to that of the previous one with the important difference that the CPB does not determine the complete allocation at each step. The substitution of one output for another is carried out by the firms depending on the common price vector for outputs announced by the CPB. Aoki (1971a) proposed a mixed planning process which combines the LAH process with Heal's (1969) process. He considers an economy with increasing returns in which there is one input such that if this is fixed, each firm faces decreasing returns with respect to all the other inputs. The CPB allocates this input to the firms in accordance with its marginal profitability and the LAH process is then used to allocate all the other resources. This process is clearly more complex since the LAH process is used at each step in which the essential input is reallocated, but it does converge to a local maximum.

Another approach to planning in economies with increasing returns is the modified LAH process. Arrow and Hurwicz (1960) showed that their process could deal with linearities and nonconvexities if the Lagrangian is suitably modified so that it becomes strictly concave and the gradient method is then applied to locate a saddlepoint of this concavified Lagrangian. There is, however, a significant difference. The modified Lagrangian expression is no longer a sum of functions each involving a different variable and it is no longer simply possible to determine demands and supplies given the prices. The CPB and firms need the entire price schedule and this makes this modified process less informationally decentralized than the original LAH process.

All the processes that we have so far considered in this section, depending as they do on first order properties of the relevant functions, cannot guarantee convergence to a global optimum. They also seem to be less informationally decentralized than processes for the classical environment. The natural question to be raised at this stage is whether it is possible to formulate a decentralized process which converges to a global optimum in an environment with increasing returns. Calsamiglia (1977) showed the answer to this question is no. He begins by making a rather important point about the interpretation of a local maximum. He provides an example of an allocation which is a local maximum but does not satisfy aggregate production efficiency. While at a local maximum it is not possible to make marginal changes to increase utility, it may be possible to increase utility simply by reorganizing production among the firms to produce more of each commodity. But, as he then proves, even in simple economies with increasing returns there does not exist a decentralized process which converges to a global optimum.

It is however, possible to construct a process which has nice convergence properties at the cost of giving up decentralization as defined in D.10. This was shown by Cremer (1977). He considers a quantity–quantity algorithm in which the CPB, as in Malinvaud (1967) and Weitzman (1970), possesses a memory and builds up successive approximations of the firms' production sets and solves the programming problem. This process is in many respects similar to Weitzman's process. Convexity of the production sets was used crucially in Weitzman's process to ensure that when the CPB constructs a new production set, by considering the announced marginal rate of substitution, it knows that no point above the corresponding hyperplane need be considered again. In the presence of increasing returns this is no longer true and in Cremer's process firms do not respond with marginal rates of substitution. The CPB only knows that if a firm responds with a feasible production plan then all production plans which are greater than it can be ruled out of further consideration. Figure 2 shows how the CPB revises its information about the firm's technology. It is assumed that the CPB knows that the optimal production plan $y^* \leq w$. It announces w as a target. If this is not feasible the firm responds

Planning, Fig. 2


boundary of the true production set
boundary of the Y^1, the CPB's first approximation
y* — true optimum
w — CPB's initial target
Commodity 2
Commodity 1
0
v^2
w
y*
y^1
v′

with some y^1 which is feasible and strictly less than w. The CPB then knows that it must now consider only points less than or equal to either v^1 or v^2.If the utility at v^2 is higher than at v^1 the CPB considers its new approximate production set to be the set of all points in Y^1 but equal to or less than v^2. Under certain boundedness conditions it can be shown that this process converges to a global optimum. Since the targets are not necessarily feasible nor is the process.

Public Goods

This section draws heavily on Tulkens (1978). We begin by considering the simple setting of an economy with a single private good y and a single public good z. There are n consumers with continuously differentiable and strictly quasi-concave utility functions $U^i(x^i)$, where $x^i = (y^i, z^i)$. The public good is produced according to the technology of the form $w = g(z)$, where w represents the private good input and g is assumed to be convex. A feasible allocation $((x^i))$ satisfies the conditions that

$$\sum_i y^i + \ w = \sum_i \omega i, \ \ z^i = z \ \text{ for all } \ i \ \text{ and } \ w = g(z).$$

Lindahl (1919) in his positive solution to the public goods problem proposed a process, the convergence properties of which were analyzed by Malinvaud (1971). The Lindahl process concerns a two consumer economy in which the public good is produced under constant marginal cost γ. Each consumer is assigned a share, θ^i in the price of the public good so that $\theta^1 + \theta^2 = \gamma$. Consumers take as given their personalized prices or unit taxes θ^i to determine their demands for the public good. The supply of the public good is made equal to the lower of these two demands and the CPB adjusts the unit taxes by raising the tax on the consumer with the higher demand and lowering it for the other. The process continues as long as the utilities of both the consumers rise. Malinvaud (1971) showed that utilities would not rise monotonically until the two demands become identical and, therefore, this process does not converge to a Lindahl equilibrium. He suggested a modification which ensures convergence to a Pareto optimal allocation (though not necessarily to a Lindahl allocation). In this modified process the CPB announces not only unit taxes but also lump-sum taxes, T^i such that $\Sigma_i T^i = 0$. Let $d^i(\theta^i, T^i)$ refer to consumer i*'s* demand for the public good and $\overline{d}$ the corresponding average demand. The CPB adjusts i's unit tax in proportion to the difference between d^i and $\overline{d}$. Supply is made equal to the average demand and T^i is adjusted to compensate i for the change in θ^i. Formally the adjustment equations are, (i) $\dot{\theta}^i = a\left[d^i - \overline{d}\right]$ for all i, (ii)$T^{\cdot i} = -(1/n)\sum_i d^i \theta^i$ for all i, where a is a positive constant. While this process converges to a Pareto optimal allocation, it is neither feasible nor monotonic.

An alternative would be to consider a process in which the CPB responds with quantities rather than prices. The Malinvaud–Drèze–de la Vallée Poussin (MDP) process, formulated by Malinvaud (1970–71) and Drèze and de la Vallée Poussin (1971), is a quantity guided process in which the CPB announces an allocation and the agents respond with rates of substitution. Starting with a feasible allocation, the firm reports its marginal cost γ and each consumer reports his or her marginal rate of substitution of the public good for the private good π^i. The adjustment takes place according to the following differential equations, (i) $\dot{z}_t = \dot{z}^i{}_t = a\left(\sum_i \pi^i{}_t - \gamma_t\right)$ for all i, (ii) $\dot{w}_t = \gamma_t \dot{z}_t$ (iii) $\dot{y}^i{}_t = \pi^i{}_t \dot{z}_t \ + \delta^i a\left(\sum_i \pi^i{}_t - \gamma_t\right)^2$ for all i, where a is a positive constant and $\delta^i \geq 0$ for all i and $\Sigma_i \delta^i = 1$.

Since the process starts at a feasible allocation, (ii) ensures that the process is feasible. It has also been shown that it converges to an allocation at which the first order conditions for optimality are satisfied, that is, the sum of the marginal rates of substitution equals the marginal cost. The MDP process is also monotonic. To see this consider

$$\dot{U}^i = U^i y^i\left(\cdot^i + \pi^i \cdot\right).$$

Using (i) and (ii) this can be rewritten as

$$\dot{U} = \ U^i y^i \delta^i a \left(\sum_i \pi^i - \gamma \right)^2 \geq 0$$

While the MDP process converges to some Pareto optimal allocation depending on the choice of the distribution profile $((\delta^i))$, Champsaur (1976) has shown that the process is neutral in the sense that given any initial allocation and any Pareto optimal allocation which is Pareto superior to this allocation, there exists a distribution profile with which the MDP process converges to the given optimum. A discrete time version of the MDP, with the same performance properties, was provided by Champsaur et al. (1977).

Malinvaud (1970–71) and Drèze and de la Vallée Poussin (1971) also extend the MDP process to an economy with many private and public goods by considering the MDP process as described above for public goods and a quantity guided process for the private goods. Another alternative, considered in Aoki (1971b), Malinvaud (1972) and Champsaur et al. (1977), is to construct a process which combines the MDP process with a price guided process for private goods. These processes, however, have to deal with a well known problem, namely one of ensuring convergence of a price guided process in an economy with many consumers without making the gross substitutability assumption.

Aoki (1971b) considers an economy with many private and public goods and many firms and consumers. He avoids the income distribution problem by specifying a social welfare function. The CPB announced prices of the private goods and quantities of the public goods. Firms maximize profits and report input demands and marginal costs for public goods. The CPB increases private goods prices according to the difference between marginal utilities and prices and the public goods levels are adjusted according to the difference between marginal utilities and marginal costs. This process is feasible, monotonic and convergent.

Malinvaud (1972) formulates a price guided process for allocating not only private goods but also public goods. The gross substitutability assumption is avoided by specifying individual incomes as proportions of aggregate income and revising them during the process (notice that Malinvaud's 1971, price guided process also made use of lump-sum transfers). This process converges locally but is neither feasible nor monotonic.

Champsaur et al. (1977) present a process which combines in a sequential way an MDP process for public goods allocation with a price guided process for private goods allocation. Given public goods' levels a price guided process is used to allocate private goods. Then keeping fixed the levels of all except one numeraire private good, the MDP process is applied to allocate public goods. This process is shown to be feasible, monotonic and convergent to some Pareto optimal allocation.

Given the difficulty in using a price guided process when there are many consumers, it is perhaps not surprising that a satisfactory process which converges to a Lindahl equilibrium has not been established, although some results are available in this direction (see Milleron 1974).

See Also

- ► Decentralization
- ► Hotelling, Harold (1895–1973)

P

Bibliography

Aoki, A. 1971a. An investment planning process for an economy with increasing returns. *Review of Economic Studies* 38: 273–280.

Aoki, A. 1971b. Two planning processes for an economy with production externalities. *International Economic Review* 12: 403–413.

Arrow, K. 1951. An extension of the basic theorems of welfare economics. In *Proceedings of the second Berkeley symposium*, ed. J. Neyman. Berkeley: University of California Press.

Arrow, K.J., and L. Hurwicz. 1960. Decentralization and computation in resource allocation. In *Essays in economics and econometrics in honor of harold hotelling*, ed. R.W. Pfouts. Chapel Hill: University of North Carolina Press; reprinted in K.J. Arrow and L. Hurwicz (1977).

Arrow, K.J., and L. Hurwicz, eds. 1977. *Studies in resource allocation processes*. Cambridge: Cambridge University Press.

Barone, E. 1908. The ministry of production in the collectivist state. In *Collectivist economic planning*, ed. F.-A. von Hayek. London: Routledge, 1935.

Calsamiglia, X. 1977. Decentralized resource allocation and increasing returns. *Journal of Economic Theory* 14: 263–283.

Champsaur, P. 1976. Neutrality of planning procedures in an economy with public goods. *Review of Economic Studies* 43: 293–300.

Champsaur, P., J.H. Drèze, and C. Henry. 1977. Stability theorems with economic applications. *Econometrica* 45: 273–294.

Cremer, J. 1977. A quantity–quantity algorithm for planning under increasing returns to scale. *Econometrica* 45: 1339–1348.

Cremer, J. 1978. A comment on 'Decentralized planning and increasing returns'. *Journal of Economic Theory* 19: 217–221.

Dantzig, G.B., and P. Wolfe. 1961. The decomposition algorithm for linear programs. *Econometrica* 29: 767–778.

Debreu, G. 1954. Valuation equilibrium and Pareto optimum. *Proceedings of the National Academy of Sciences* 40: 588–592.

Debreu, G. 1959. *Theory of value: An axiomatic analysis of economic equilibrium*, Cowles Foundation Monograph no. 17. New York: Wiley.

Drèze, J.H., and D. de la Vallée Poussin. 1971. A tâtonnement process for public goods. *Review of Economic Studies* 37: 133–150.

Heal, G.M. 1969. Planning without prices. *Review of Economic Studies* 36: 346–362.

Heal, G.M. 1971. Planning, prices and increasing returns. *Review of Economic Studies* 38: 281–294.

Heal, G.M. 1973. *The theory of economic planning*. Amsterdam: North-Holland.

Hori, H. 1975. The structure of the equilibrium points of Heal's process. *Review of Economic Studies* 42: 457–467.

Hurwicz, L. 1960. Optimality and informational efficiency in resource allocation processes. In *Mathematical methods in social sciences*, ed. K.J. Arrow, S. Karlin and P. Suppes. Stanford: Stanford University Press. Reprinted in Arrow and Hurwicz (1977).

Hurwicz, L. 1969. On the concept and possibility of informational decentralization. *American Economic Review* 59: 513–534.

Hurwicz, L. 1973. The design of resource allocation mechanisms. *American Economic Review* 58: 1–30; reprinted in Arrow and Hurwicz (1977).

Kornai, J., and T. Liptak. 1965. Two-level planning. *Econometrica* 33: 141–169.

Lange, O. 1936–7. On the economic theory of socialism. *Review of economic studies*. Reprinted in *On the economic theory of socialism*, ed. B. Lippincott, Minneapolis: University of Minnesota Press, 1938.

Lindahl, E. 1919. Just taxation – A positive solution. In *Classics in the theory of public finance*, ed. R.A. Musgrave and A. Peacock. London: Macmillan, 1958.

Malinvaud, E. 1967. Decentralized procedures for planning. In *Activity analysis in the theory of growth and planning*, ed. M.O.L. Bacharach and E. Malinvaud. London: Macmillan.

Malinvaud, E. 1970–71. Procedures pour la determination d'un programme de consommation collective. *European Economic Review* 2: 187–217.

Malinvaud, E. 1971. A planning approach to the public goods problem. *Swedish Journal of Economics* 73: 96–112.

Malinvaud, E. 1972. Prices for individual consumption, quantity indicators for collective consumption. *Review of Economic Studies* 34: 385–406.

Marglin, S.A. 1969. Information in price and command systems of planning. In *Public economics*, ed. J. Margolis and H.K. Guitton. London: Macmillan.

Milleron, J.C. 1974. *Procedures to a Lindahl equilibrium corresponding to a given distribution of income*. Paper presented at the European Meeting of the Econometric Society in Grenoble.

Pareto, V. 1897. *Cours d'économie politique*. Vol. 1. Lausanne: Librarie de l'Université.

Samuelson, P.A. 1949. Market mechanisms and maximization. In *Collected scientific papers of Paul A. Samuelson*, ed. J.E. Stiglitz. Cambridge, MA: MIT Press, 1966.

Taylor, F.M. 1929. The guidance of production in a socialist state. *American Economic Review* 19: 1–8. Reprinted in On the *economic theory of socialism*, ed. B. Lippincott. Minneapolis: University of Minnesota Press, 1938.

Tulkens, H. 1978. Dynamic processes for public goods: An institution-oriented survey. *Journal of Public Economics* 9: 163–201.

Uzawa, H. 1958. Iterative methods for concave programming. In *Studies in linear and non-linear programming*, ed. K.J. Arrow, L. Hurwicz, and H. Uzawa. Stanford: Stanford University Press.

von Hayek, F.A. 1935. The present state of the debate. In *Collectivist economic planning*, ed. F.A. von Hayek. London: Routledge.

von Mises, L.. 1922. *Socialism*. London: Jonathan Cape, 1936.

Weitzman, M. 1970. Iterative multi-level planning with production targets. *Econometrica* 38: 50–65.

Plant, Arnold (1898–1978)

Ronald H. Coase

Sir Arnold Plant, Professor of Commerce (with special reference to Business Administration) at the London School of Economics, was born in 1898 in Hoxton, in East London, the son of a municipal librarian. He died in 1978. He served on numerous government committees, particularly in the years after World War II, and was knighted in 1947. Most of his important contributions to economics were reprinted in *Selected Economic Essays and Addresses* (1974).

On leaving school he joined a mechanical engineering organization and in 1920 became manager of the Steam Fittings Company. Advised by William Piercy (later Lord Piercy) that he ought to learn something about management before doing much more of it, he enrolled at the LSE, where he studied for the BCom as an external student and the BSc(Econ.) as an internal student, specializing in modern economic history. He was awarded the BCom in 1922 and the BSc(Econ.) in 1923. The teacher who most influenced him was Edwin Cannan, the Professor of Political Economy, whose views and commonsense approach to economic analysis and economic policy were to be reflected in Plant's own work.

In 1923, Plant was appointed Professor of Commerce at the University of Cape Town. The only one of his writings in South Africa reprinted in the 1974 volume was an essay published in 1927 dealing with the economic relations of the races. It was a trenchant attack on the South African government's policy of separation of the races. This policy, Plant argued, arose out of a desire to stifle competition from the native peoples and was economically injurious to South Africa. He advocated the provision of educational opportunities and other measures designed to bring the natives into Western society. While at Cape Town he gathered materials later used in the chapter on Economic Development which he wrote for the volume on South Africa of the *Cambridge History of the British Empire* (1936).

In 1930 Plant became Professor of Commerce at LSE. In addition to his lectures to the BCom students, he also taught in the postgraduate Department of Business Administration, of which he later became Head. His analytical system was unsophisticated but powerful. He thought of the economic system as essentially competitive, with monopoly transitory and relatively unimportant. The State had a role in providing law and order but State intervention commonly promoted monopoly and, as in South Africa, was designed to promote the interests of groups with political power. He was especially interested in attempting to understand the reasons which led to the adoption of various business practices.

Inspired by David Hume's treatment of the subject, he developed an interest in property and the economic function of ownership. This led Plant to write in the early 1930s two articles, one on patents for inventions and the other on copyright in books (both reprinted in the 1974 volume). These articles rank as his major achievement in economics. He questioned the need for property rights in patents and copyright. They did not arise out of scarcity but created scarcity by establishing a monopoly. He pointed out that British authors had received handsome incomes in America although they possessed no copyright there while much invention goes on even though the resulting improvements are not patentable. If the existing law was to be retained, he suggested that modifications should be made such as making Licenses of Right the normal practice in the case of patents.

Selected Works

1936. (ed.) *The Cambridge history of the British Empire*. Vol. 8: *South Africa, Rhodesia and the Protectorates: Economic development, 1795–1921*. Cambridge: Cambridge University Press.

1974. *Selected economic essays and addresses*. London: Routledge.

Plantations

Adrian Graves

The economic, social and political importance of plantations in many regions, the longevity and ubiquity of the institution, its association with slavery and other forms of bonded labour and with colonialism, has given rise to an extensive and rich literature which spans many scholarly disciplines including history, sociology, politics, psychology, anthropology, archaeology and geography. Economists and political economists have been preoccupied with explaining the origins of plantations and evaluating their social and economic effects, both locally and in the broader context of the world economy. A survey of the intellectual origins and thrust of the most recent economic literature, however, illustrates the immense difficulties of theorizing the plantation. The failure to derive universally applicable definitions of the plantation and of the plantation economy lies at the heart of the problem.

The meaning of 'plantation' has changed markedly over time. Originally, it referred to a plot of ground set with plants or trees. With the onset of British overseas expansion, plantation officially designated a group of settlers or their political units, hence 'Ulster Plantation' or the 'Caribbean Plantations', but this usage was eventually replaced by 'colony' and the use of 'plantation' became restricted to farms or landed estates. In this sense, the word has been applied especially in tropical or subtropical countries though units of agricultural production in temperate or non-tropical zones, such as in parts of Europe and the Middle East, have also been referred to as plantations.

The range of plantation crops is also extremely diverse, including sugar, coffee, tobacco, tea, cocoa, bananas and other tropical or sub-tropical fruits, chewing gum (chicle), rice, tree spices, such as nutmeg, cloves, cardamon, mace, vanilla, cinnamon, garden crops like ginger and pepper and industrial raw materials such as cotton, indigo, copra, oil palm, sisal, cinchona and rubber. Although plantations are frequently typified as monocultural institutions, many plantation products were grown either as subsidiary crops or they were combined with the cultivation of cereals, temperate zone fruits such as citrus, market garden crops and even with livestock production.

Most plantations combined an agricultural with an industrial process, though they were not of necessity bifurcated institutions. The scale and sophistication of the technological forms on plantations and associated infrastructures, as well as the structures of ownership and control of plantations have varied markedly according to time and location. Plantations have also seen many modal transformations, being based on slavery, a variety of feudal forms, peonage systems, long contract migratory or indentured labour and free wage labour. There are many examples of plantations, for instance in Latin America, that operated on a mixture of labour forms, in economies which articulated around a variety of modes of production. The extraordinary diversity of geographical location, crops, sources of labour, ownership or control and technological forms over time have created major difficulties for the scientific definition of the plantation.

All definitions of the plantation attempt to differentiate it from other agricultural or agro/industrial institutions, frequently, by very general characteristics such as climate, crop type or specialization, export orientation, spatial size, number of employees, or by its system of power or authority structure. Some writers lay particular stress on labour force characteristics, its degree of bondage, skill levels, the tendency of work to be organized cooperatively around gangs, the stability or permanence of the workforce, and cultural and ethnic or racial criteria. For others, factor proportions are paramount although the emphasis on specific factor ratios varies widely in the literature. Whereas a number of writers emphasize the capital intensive nature of plantation production (such as Paige's high capital/labour approach: Paige 1975), others (e.g Stinchcombe 1961) define plantations as peculiarly land intensive units of production (see Pryor 1982, pp. 289–91).

Since most definitions are developed to service the analysis of a particular economy, region or timespan, the enormous degree of variability in plantation production has led to a plethora of definitions stressing markedly different criteria which are frequently contradictory. This is particularly evident, as we have already noted, in the literature which typifies plantations according to factor proportions, or that which lays stress on the level of work skill, which is particularly low according to some writers (e.g. Baldwin 1956; Stinchcombe 1961) and especially high according to others (e.g. Wolf and Mintz 1957). Precise definitions of the plantation inevitably exclude institutions which might justifiably be considered as such: the stress on the unfree nature of the plantation labour force, for example, rules out consideration of plantations based on wage labour. The most general definitions, however, inevitably incorporate a wide range of other agricultural institutions, including production units as varied as Roman latifundia, large estates in Byzantine Egypt, feudal estates, colonial agricultural missions, cooperative estates, state farms, modern capital or labour-intensive sugar or cotton estates, ranches or pastoral stations, and large corporate farms. Needless to say, the scholarly difficulties of defining plantations are reflected in the attempts to theorize the broader concept of the plantation economy.

The theory of the plantation economy has a long and rich intellectual pedigree, drawing upon classical and marxist traditions. Its classical intellectual origins can be traced back to the debates on land/labour ratios in the early 19th century which involved Ricardo, Wakefield, Torrens, Merivale and John Stuart Mill. Merivale, in particular, stressed the dominant role of the plantation in the ex-slave economies of the West Indies. It was left to the Dutch scholar H.J. Nieboer to expand the scope of the debate by transforming the undifferentiated notion of the influence of the plantations in the tropical economies into a general theory. He characterized differing types of colonial society according to the theory of open and closed resource systems, identifying the initial absence of permanent settlement as a pre-condition of plantation production, the necessity of formal compulsion of the labour force, and the subsequent engrossment of the optimum available land as the hallmarks of plantation production and subsequently of plantation economies and societies. In the 1930s and 1940s, Edgar T. Thompson incorporated the concepts of open and closed resource systems into a theory of social change in plantation society, through which he attempted to identify disintegrating forces inherent in the plantation which also emerged as important factors in the wider economy and society.

The influence of Nieboer and Thompson was extended by studies of wage and slave plantations undertaken in the 1950s and 1960s. The work of Mintz and Wolf stands out in this respect, especially in an important article on plantation society in Central America and the Caribbean which distinguished haciendas from plantations and developed the notion of old and new style plantations (Wolfe and Mintz 1956). Whereas the former were precapitalist with surpluses directed at conspicuous consumption, the latter were typified as capitalist enterprises driven by the process of surplus extraction to service capital accumulation. This work opened up associated discussions in marxist theory on the nature of a mode of production in which distinctions between slave plantations and capitalist institutions continue to play an important part. But it was the work in the late 1960s and early 1970s, of a group of Caribbean social scientists known as the New World Group, most prominently Lloyd Best and George Beckford, that attempted to integrate the classical literature on slave plantations with the then emerging marxist debate on underdevelopment as a means to analyse post slave plantation production. This work has been extremely influential in the most recent literature on plantation production and warrants closer examination.

Best’s contribution has been to try and develop a universally applicable model of a ‘pure plantation economy’. In so doing, he drew heavily upon the intellectual heritage of Nieboer and Thompson, incorporating also Erving Goffman’s notion of the) ‘total institution’:

Where land is free to be used for subsistence production the recruitment of labour exclusively for export production imposes a need for ‘total economic institutions’ so as to encompass the active existence of the workforce. The plantation which admits virtually no distinction between organisation and society, and chattel slavery which deprived workers of any civil rights including the right to property, together furnish an ideal framework (Best 1968, p. 287).

This conceptualization was subsequently elaborated by Beckford (1972) to include a vigorous critique of the dual economy and, most significantly, to demonstrate the meagre spread effects of modern plantation production.

Beckford distinguished between the mainly temperate) 'colonies of settlement' (Australia, New Zealand, Canada and the United States of America) and tropical 'colonies of exploitation', observing that the pattern of agricultural production which emerged in the two types of colonies was significantly different. Generally speaking, Beckford's attempts to develop a theory of modern plantation production rests upon this basic distinction, (as does the more recent work of de Silva 1982) with plantations being identified firmly with the tropical colonies of exploitation. Although he specified two exceptional types of plantation economy the plantation sub-economy (USA, north–eastern Brazil, Honduras, Guatemala, Costa Rica and Panama) and the enclave plantation economy (Liberia, Kenya, Rhodesia and South Africa) Beckford based his argument about the meagre spread effects of plantations on a general model of the plantation economy, namely,

> Those countries of the world where the internal and external dimensions of the plantation system dominate the country's economic and social and political structures and its relations with the rest of the world . . . wherever several plantations have come to engross most of the arable farm land in a particular country which is predominantly agricultural, that country can be described as a plantation economy or society and its social and economic structure and external relations will be similar to those described for the plantation system (Beckford 1972, p. 12).

Whilst he recognized that the potential for efficient resource use within the firm or particular plantation, Beckford argued that plantation agriculture was essentially unproductive, and that plantations exercised a pervasive economic, social and political influence over their areas which reinforced and perpetuated the underdevelopment of those economies. The fact that these local characteristics occurred within the wider context of dependent, exploitative, metropolitan/periphery financial and trading relationships is an important component of Beckford's argument concerning the meagre spread effects of plantation production. His general conclusion was that) 'regardless of the type of plantation that predominates in any given situation, the result is always the same – a persistent tendency towards underdevelopment (Beckford 1972, p. 213). It is important to appreciate that Beckford's typology of plantation production was meant to be appropriate for all major areas of plantation production regardless of whether or not the institution had originated in the slave based mercantilist empires of the New World.

The most recent studies of plantation economies have built upon the important work of the New World Group. Mandle in particular owes a great debt to Beckford, but he also incorporates other long-standing assumptions in the literature on plantation production. Thus he puts great stress upon the innate inefficiency and inflexibility of plantation production. While this was due to a number of factors, the characteristically coerced and cheap plantation labour force was the 'key' to low productivity, because it provided planters with little incentive to innovate at the critical level of cultivation technology or to escape from their dependence upon international markets. It was partly through his stress on labour force characteristics, that Mandle underlined the importance of going beyond the analysis of the plantation merely as a productive unit (compared, for example, with the approach of the frequently quoted William O. Jones: see Jones 1968) to embrace the distinctive kinds of social and production relations which derive from the plantation structure. On this basis, he distinguished from the growth-oriented capitalist mode of production, his notion of the) 'plantation mode of production' by which he meant the) 'growth inhibiting social structure' typical of plantation economies. Four attributes characterize Mandle's plantation mode of production; large-scale agriculture dominates the society; the domestic labour supply is inadequate to meet the labour demands of the agricultural sector; labour is mobilized and allocated by non-market mechanisms (coercion) which in turn define the nature of class relations in the society; and these class relations are reinforced by a distinctive culture (Mandle 1982, pp. 37–8). The pivotal idea in Mandle's work, that the plantation mode of production inherently constrains its own technological advance and therefore the broader development of the forces of production in the plantation economy, has been supported in important studies of Louisiana agriculture by Ferleger (e.g. Ferleger 1984) and although it does not necessarily imply Mandle's model, the phrase 'plantation mode of production' is assuming a wider currency in the literature.

Although it has contributed considerably to our understanding of the institution's economic, political and social impact, the recent literature illustrates the formidable difficulties of developing a dynamic economic theory of the plantation. For one thing, it confuses model building with theorizing. Theory in history must incorporate process. Unquestionably, the recent theorists of the plantation recognize that the current state of the economies they address has historical roots. In the final analysis, however, the 'theories' of Best, Beckford and Mandle, go little beyond listing a set of distinctive, generalized characteristics of the plantation and of the plantation economy. To that extent, their typologies are static and therefore ahistorical. At another level, the primacy they place on the nature of the plantation itself and its function in the wider economy and society, effectively denies a role to human agency in the history of plantation economies.

To the extent that the literature incorporates the notion of underdevelopment, it is subject to the same sorts of criticisms which have more recently been marshalled with some force against André Gunder Frank and others (including Immanuel Wallerstein) by writers from both the marxist and neoclassical perspectives. In particular, Beckford's concern to identify the domination of the world market and the demands of capital reproduction as the arbiter of the rate of capital accumulation in plantation economies deprives his typology of the plantation of any laws of motion. Do plantation economies merely vegetate 'on the periphery of an industrializing Europe like a vast reservoir of labour-power periodically called into action by the spasmodic actions of metropolitan capital,' like the underdeveloped ex-colonies of Frank and Laclau's Latin America? (Banaji 1977, p. 14).

As to the 'plantation mode of production', other writers have noted the analytical difficulties of the more general concept of mode of production. Even within Marx, the application of the concept is confusing and even contradictory (does it mean the 'labour process' or an 'epoch of production': Banaji 1977, pp. 4–5). But the problems of that idea are compounded substantially when it is ahistorically applied to forms of production which cover a wide chronological range, incorporating changing, even co-existing, different patterns of production relations, as McEachern (1976) has demonstrated in his criticisms of Alavi's 'colonial mode of production'. The same sorts of criticisms can be addressed to Mandle's apparently timeless notion of the 'plantation mode of production'.

Beyond these broad points, a major weakness of the recent literature arises from its tendency to generalize about the impact of (exclusively sugar cane) plantations on the basis of a Caribbean/American paradigm. In respect of that region alone, other research casts doubt on the supposed inflexibility and inertia of plantation production even under slavery (Fogel and Engerman 1974; Drescher 1977; Ward 1985) but also under systems of indentured labour (Saha 1970; Beechert 1987). Moreover, the characteristics and performance of sugar plantations in other regions, both in micro and macro terms, do not fit the Beckford/Mandle typologies. In 19th-century Queensland, Natal and Hawaii, for example, plantations did not of necessity dominate their areas and they proved to be not only highly flexible and dynamic institutions experiencing revolutionary changes in production over relatively short time periods, they stimulated rather than retarded local economic development, as have plantations more recently in Kenya (Graves and Richardson 1980; Pryor 1982; Beechert 1987). Recent studies of colonial Latin America show that dynamic growth and flexible response was also a feature of plantation production in underdeveloped economies as well, not only in respect of cane sugar production but also in coffee, an important crop which is all but ignored by the theorists of the plantation economy (Duncan and Rutledge 1977).

There is no doubt that much of the dynamism of plantations, especially in the late 19th century, was due to a revolution in the processing technology of the major plantation crops. This performance, contrary to the claims of the Mandle model, was also attributable on sugar plantations to significant innovations in the plantation cultivation process under production imperatives and structures which were unambiguously capitalist. While it is true that cane harvesting equipment was not successfully employed until the mid-20th century, the immense technical difficulties of that sort of machinery cannot be underestimated. Important innovations in cultivation technology were not confined to the more developed economies which supported plantation production. Java, for example, boasted a long history of agricultural research and development in the 19th and early 20th centuries which saw amongst other achievements, the emergence of the famous P.O.J.100 cane variety which transformed cane cultivation on an international scale. Less spectacular but no less significant innovations in the cultivation process were frequently introduced in many plantation economies by workers to raise their productivity under piece rate regimes, such as the redesigning of cane knives to suit local conditions.

Despite its weaknesses, the best recent literature on plantation production has attempted the important and urgent task of identifying the distinct economic rhythms and movements of modern colonial plantation production. Whilst the literature owes a rich debt to the classical formulations on plantations, it has also accommodated more recent theoretical approaches to colonialism and metropolitan–periphery relations. It is evident, however, that it is extremely difficult to theorize the plantation. Whilst the immense problems of defining a plantation suggest that its use as a descriptive term of agricultural organization its itself problematic, the scientific rigour of 'plantation' has very severe limitations. It appears to be unable to make analytic distinctions between agricultural production of tropical produce under the markedly different labour forms of slavery, feudal labour service, indentured labour, peonage, short-contract immigrant workers, or free labour. Nor does it distinguish between monopolistic and competitive land conditions. The revolutionary impact of changes in technology has not been sufficiently well accommodated.

Studies of plantations outside the Caribbean and the Americas indicate that the 'plantation economy' cannot be understood merely in terms of the logic and form of the plantation as a productive unit or of the plantation sector, but in the various combinations of relations of production which characterize the particular economies within which plantations operate. The explanation for the character, persistence or transformation of plantations therefore, must go beyond the discrete analysis of the institution itself and be sought more explicitly in the demands of capital accumulation under specific and changing conditions of capital markets and land ownership, labour availability and productivity, the changing technologies of cultivation and industrial processes and the consumption and distribution structures of plantation products. Only then will the laws of motion of the plantation become apparent, as will the forces tending to undermine or conserve this form of production. There is no doubt that the considerable gaps in our knowledge about plantations, despite an extraordinarily rich scholarly literature on the institution, is due to the paucity of rigorous empirical studies which address the wider conceptual issues. While that remains the case, economic theories of the plantation and of the so-called plantation economies will continue to remain unsatisfactory.

See Also

- ► Colonialism
- ► Slavery

Bibliography

Alavi, H. 1975. India and the colonial mode of production. In *The socialist register, 1975*, ed. R. Miliband and J. Saville. London: Mertin Press.

Baldwin, R.E. 1956. Patterns of development in newly settled regions. *The Manchester School* 24: 161–179.

Banaji, J. 1977. Modes of production in a materialist conception of history. *Capital and Class* 3: 1–44.

Beckford, G.L. 1972. *Persistent poverty: Underdevelopment in plantation economies of the third World*. New York: Oxford University Press.

Beckford, G.L. 1969. The economics of agricultural resource use and development in plantation economies. *Social and Economic Studies* 18(4): 321–347.

Beechert, E. 1986. Technology and the plantation. In *Proceedings of the second World plantation conference*, ed. S. Eakin and J. Traver. Shreveport/Baton Rouge: Louisiana State University.

Benn, D.M. 1974. The theory of plantation economy and society: A methodological critique. *The Journal of Commonwealth and Comparative Politics* 12(3): 249–260.

Best, L. 1968. A model of a pure plantation economy. *Social and Economic Studies* 17(3): 283–316.

Courtenay, P.P. 1965. *Plantation agriculture*. New York: Praeger.

de Silva, S.B.D. 1982. *The political economy of underdevelopment*. London: Routledge & Kegan Paul.

Drescher, S. 1977. *Econocide: British slavery in the era of abolition*. Pittsburgh: University of Philadelphia Press.

Duncan, K., and I. Rutledge. 1977. *Land and labour in Latin America: essays on the development of Agrarian capitalism in the nineteenth and twentieth centuries*. London: Cambridge University Press.

Ferleger, L. 1984. Self-sufficiency and rural life on Southern farms. *Agricultural History* 58(3): 314–329.

Fogel, R.W. and Engerman, S.L. 1974. *Time on the cross*. Vol. 1, The economics of American slavery; Vol. 2, Evidence and methods – A supplement. Boston: Little, Brown.

Frank, A.G. 1967. *Capitalism and underdevelopment in Latin America*. New York: Monthly Review Press.

Genovese, E.D. 1965. *The political economy of slavery: Studies in the economy of the slave south*. New York: Pantheon.

Graves, A., and P.G.L. Richardson. 1980. Plantations in the political economy of colonial sugar production: Natal and Queensland, 1860–1914. *Journal of Southern African Studies* 6(2): 214–222.

Higman, B.W. 1969. Plantations and typological problems in geography. *Australian Geographer* 11(2): 192–203.

Jones, W.O. 1968. Plantations. In *International encyclopaedia of the social sciences*, ed. D.L. Sills. New York: Macmillan.

Mandle, J.R. 1982. *Patterns of Caribbean development*. New York: Gordon and Breach Science Publishers.

McBride, G.M. 1934. Plantation. In *Encyclopaedia of the social sciences*. New York: Macmillan.

McEachern, D. 1976. The mode of production in India. *Journal of Contemporary Asia* 6(4): 444–457.

Nieboer, H.J. 1900. *Slavery as an industrial system: Ethnological researches*. The Hague: Martinus Nijhoff.

Paige, J. 1975. *Agrarian revolution: Social movements and export agriculture in the underdeveloped World*. New York: Free Press.

Pan American Union. 1959. *Plantation systems of the new World*. Vol. 7, Social science monographs. Washington.

Pryor, F.L. 1982. The plantation economy as an economic system. *Journal of Comparative Economics* 6(3): 288–317.

Saha, P. 1970. *Emigration of Indian labour (1834–1900)*. Delhi: People's Publishing House.

Stinchcombe, A.L. 1961. Agricultural enterprise and rural class relations. *American Journal of Sociology* 67: 165–176.

Thompson, E.T. 1975. *Plantation society, race relations and the south: The regimentation of population*. Durham: Duke University Press.

Thompson, E.T. 1983. *The plantation: An international bibliography*. Boston: G.K. Hall.

Wallerstein, I. 1974. *The modern World-system: Capitalist agriculture and the origins of the European World-economy in the sixteenth century*. New York: Academic Press.

Ward, J.R. 1985. *Poverty and progress in the Caribbean, 1800–1960*. London: Macmillan.

Weber, M. 1927. The plantation. In *General and Economic History*, ed. M. Weber. Trans. F.H. Knight. New York: Greenburg.

Wolf, E.R., and S.R. Mintz. 1957. Haciendas and plantations in middle America and the Antilles. *Social and Economic Studies* 6(3): 380–412.

Playfair, William (1759–1823)

H. E. Egerton and F. Y. Edgeworth

Playfair attempted in his youth with little success to combine the positions of inventor and tradesman. He went to Paris, and in 1789 became agent to an American Land Company, the operations of which were disastrous to those sent out. On returning to London he opened a 'Security' Bank, which, however, soon collapsed. After Waterloo he returned to Paris as editor of *Galignani's Messenger*, but had to leave France to avoid imprisonment on a judgement in an action for libel. His publications were very numerous; many were directed against the French, and he advocated the issue of forged assignats. In the *Gentleman's Magazine* (1823, pt. i. 564) is an imperfect list of forty-one pamphlets and books, among which are *A General View of the Actual Force and Resources of France* (1793); *Better Prospects to the Merchants and Manufacturers of Great Britain* (1793); *Letter to Sir Wm. Pulteney on the establishment of another Public Bank in London* (1797); *Statistical Tables, from the German of Boetticher* (1800); *Statistical Account of the US from the French* (1807).

He published anonymously in 1785 *The Increase of Manufactures …*, a proposal to establish a fund for lending sums of money at an interest suited to the circumstances of each case. In 1786 appeared *The Commercial and Political Atlas* (brought up to date in two successive editions 1787 and 1801), remarkable for the application of the graphical method to the statistics of finance. The method is thus introduced:

Suppose the money that we pay in any one year for the expense of the navy were in guineas, and that these guineas were laid down upon a large table in a straight line and touching each other, and those paid next year were laid down in another straight line, and the same continued for a number of years, these lines would be of different lengths as there were fewer or more guineas; and they would make a shape, the dimensions of which would agree exactly with the amount of the sum (*Atlas*, 1st edition; the illustration is varied in subsequent versions).

By this method 'as much information may be obtained in five minutes as would require whole days to imprint on the memory ... by a table of figures'. Thus ordinates at points on a horizontal line represent the amount of exports and of imports at each epoch; the difference between them – forming a stream of varying width – represents the balance of trade. That Playfair should give prominence to this conception is remarkable, as his observations on our trade with France evidence a just sense of the mutual interests of the parties to international trade.

The *Real Statement of the Finances and Resources of Great Britain* (1796) contains some good remarks on the depreciation of money: 'If money should decrease in value faster than the debts increase, then the burdens of the people, though nominally augmenting, may be actually diminishing.' The rudimentary idea of an index number may be noticed in the Appendix, p. 29.

In the *Inquiry into the ... Causes of the Decline and Fall of ... Nations*, which appeared in 1805, Playfair pretends to apply his method to ancient history. In the preface he acknowledges obligation to his brother Professor John Playfair for the idea of the new method. In the same year (1805) Playfair published an edition of *Wealth of Nations*, which contains some acute criticisms; for instance, on Adam Smith's doctrine that 'the more a man pays for the tax the less he can afford to pay for the rent' [of a house] (*Wealth of Nations*, V, ch. ii); and on Sir Matthew Decker's observation approved by Adam Smith that) 'certain taxes are in the price of certain goods, sometimes repeated and accumulated four or five times' (ibid.). These are supplementary chapters on occurrences in finance subsequent to Adam Smith's time, and on the French 'Economists'. Playfair evinces some acumen as an economist as well as some originality as a statistician.

Pleasure and Pain

F. Y. Edgeworth

Pleasure and pain are the only motives taken account of in political economy in so far as 'it makes entire abstraction of every other passion or motive but the desire for wealth; except those which may be regarded as perpetually antagonizing principles to the desire of wealth, namely, aversion to labour and desire of the present enjoyment of costly indulgences' (Mill, *Unsettled Questions*, p. 138). This abstraction, legitimate within limits, is liable to be strained too far in several directions.

(1) Because economic action is ascribed to utility, it is not to be taken for granted that, as utilitarians have postulated, all action is motived by pleasure. For perhaps 'all that mathematical economics need to assume is that a material quantity of goods will be in a certain proportion to a greater or less strength of motive; whether the motive be taken as "pleasure" or not is not essential' (Bonar, *Philosophy and Political Economy*, p. 224; cf. Sidgwick, *Political Economy*, Bk. i, ch. ii, § 2 note; Marshall, *Principles of Economics*, 3rd edn, pp. 77, 78, note; and *Economic Journal*, vol. iii, p. 388). However, when equilibrium is regarded as the position of greatest advantage to all concerned

(cf. Marshall, *Principles of Economics*, 3rd edn, 526–7, and note xiv), the mechanical analogue being not so much the equality of forces (conceived by Jevons in his analogy of the lever, *Theory*, ch. iv) as the maximum of energy (indicated by Irving Fisher in his *Mathematical Investigations*), there is taken for granted the possibility of summing up pleasures which some opponents of utilitarianism have refused to grant.

(2) For the most abstract part of economics, the theory of exchange, it need not be postulated that each party acts from self-interest, but only that he is not actuated by regard for the interest of the other parties, those with whom he competes or bargains. The efforts and sacrifices which are required to supply markets – including the labour market and the loan market – are often incurred for the sake of one's family rather than oneself. The action of the family affections 'has always been fully reckoned with by economists, especially in relation to the distribution of the family income between its various members, the expenses of preparing children for their future career, and the accumulation of wealth to be enjoyed after the death of him by whom it has been earned' (*Principles of Economics*, Bk. i, ch. v, § 7, 3rd edn).

(3) The limits within which self-interested action must be postulated may be even narrower than those indicated in the last paragraph. What is postulated is that action should be regular and therefore calculable, rather than that it should be self-interested (*Principles of Economics*, Bk. i, ch. v). 'The range of economic measurement may gradually extend to much philanthropic action.'

Bibliography

Bonar, J. 1893. *Philosophy and political economy.* London: S. Sonnenschein & Co.

Cournot, A.A. 1838. *Researches into the mathematical principles of the theory of wealth.* Trans. by N.T. Bacon with a bibliography of mathematical economics by I. Fisher. New York/London: Macmillan, 1897.

Fisher, I. 1892. Mathematical investigations in the theory of value and prices. *Transactions of the Connecticut Academy of Arts and Sciences* 9: 1–124.

Jevons, W.S. 1871. *The theory of political economy.* London: Macmillan.

Marshall, A. 1890. *Principles of economics*, 3rd ed. London: Macmillan, 1895.

Marshall, A. 1893. Meeting of the British Economics Association. *Economic Journal* 3(3): 377–390.

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

Mill, J.S. 1844b. *Essays on economics and society*, 2 vols. London.

Sidgwick, H. 1883. *Principles of political economy.* London: Macmillan.

Stephen, L. 1900. *The English utilitarians*, 3 vols. London: Macmillan.

Plekhanov, Georgii Valentinovich (1856–1918)

M. Falkus

Keywords

Capitalism; Lenin, V. I.; Marxism; Plekhanov, G. V.; Socialism

JEL Classifications

B31

Plekhanov was a major figure in the development of Marxist economic and political philosophy during the late 19th century. His importance springs from four principal sources. He was the first Russian intellectual to apply Marxist theory to Russian conditions. In so doing, he undermined the intellectual foundations of the Populists (*Narodniki*) and showed the relevance of Marxist economic determinism to Russia. Secondly, he exerted a profound influence upon the Russian revolutionary intelligentsia, persuading many of them to abandon Populism in favour of Marxism. Plekhanov was one of the founders of the Marxist Russian Social Democratic Party. Thirdly, the originality and perception shown in Plekhanov's own voluminous and wide-ranging writings show him to be an outstanding Marxist theoretician. Finally, the approval given Plekhanov's writings by Marx and, especially, Lenin (despite their later disagreements) has assured Plekhanov of an honoured place in Soviet histories of the development of socialist philosophy. Indeed, Plekhanov was one of the two figures whose writings were specifically acknowledged by Lenin as leading to his own conversion to Marxism; the other was Marx.

Plekhanov was born on 29 November 1856 in the village of Gudalovka in what was then the province of Tambov (Lipetsk Oblast). He was the son of a wealthy nobleman and attended military college in Voronezh and the Konstantin Cadets' College in St Petersburg in 1873–4 before entering the St Petersburg Institute of Mines. Here he became influenced by the revolutionary movements of the time and was eventually expelled in 1876 for his part in such activities. In 1875 he had joined the Narodniki and in the following year he joined the newly formed *Zemlya i Volya* (Land and Liberty) *Narodnik* organization – Russia's first political party. This group believed that Russia's future lay with the peasant masses, and that the peasants should be given land. Plekhanov soon became one of the leading *Narodnik writers* and activists, and took part in the 'going to the people' movement. He also gave a speech at a major demonstration organized in 1876 by *Zemlya i Volya* in front of St Petersburg's Kazan Cathedral.

In 1879 the *Narodnik* movement split, the majority faction advocating the use of terrorist tactics. Plekhanov favoured a more moderate approach, and together with a small group of other leading *narodniks* (including Pavel Axelrod and Leo Deutsch) formed the non-violent *Cherny Peredel* movement (Black Repartition – that is, the movement wanted repartition of the fertile Black Soil lands to the peasantry).

In January 1880 Plekhanov emigrated to Europe to escape persecution from the tsarist authorities. He remained in exile until 1917, living in Switzerland, France, Italy and elsewhere, travelling widely throughout the continent. In western Europe he made contact with numerous other Russian revolutionary exiles and also became deeply interested by Marxist thought. From about 1882 he became a fervent advocate of Marxism, and in his writings he now sought to establish the relevance of Marxism to Russian conditions and to undermine the intellectual foundations of Russian Populism. In 1883 Plekhanov founded in Geneva Russia's first Marxist Social Democratic organization, the Liberation of Labour. The group translated into Russian and published many works by Marx and Engels, Plekhanov himself translating the *Communist Manifesto*.

During the 1880s and 1890s Plekhanov wrote his most influential works, denouncing not only the Populists but the Legal Marxists and the Economists (Marxist factions which developed after 1895), and he put forward his own interpretation of the path towards socialism which Russia was to follow. The root of his philosophy was in what he termed 'scientific' historical materialism, exposing the *narodniks* as 'unscientific'. In Plekhanov's view, revolution could not succeed unless it has the support of the class-conscious masses. Revolution could not come from the agrarian peasantry, and must come from the urban proletariat. As he argued in *Socialism and the Political Struggle* (1883) and *Our Differences* (1885a), the utopian socialists (Blanquists) were mistaken in their reliance on intellectual conspiracy alone: revolution could succeed only as a result of a class struggle emanating from the working classes. It therefore became important for Plekhanov to demonstrate that Russia's path towards socialism could not come, as the *narodniki* argued, from the village-based commune (*mir*) and the peasantry. Capitalism in Russia was a necessary phase of historical development and was not 'accidental' or 'non-Russian'. Indeed, in Russia of the 1880s capitalism was already a reality.

To be sure, Plekhanov's theories contained many obscurities and contradictions.

Fundamental were the dichotomies between economic determinism and the role of the revolutionary, and also between the reliance on the class-conscious urban masses and the evident industrial backwardness of Russia. Plekhanov 'solved' the problems, albeit unsatisfactorily, by arguing that the Russian revolution could be accelerated by the role of the revolutionary intelligentsia, whose activities were to compensate for the lack of a middle class. He wrote in *Our Differences*: ‘Our capitalism will fade without ever having flowered.’

Particularly influential was Plekhanov’s *The Development of the Monistic View of History*, which was brought to Russia by the Marxist publisher Potresov in 1894. Here Plekhanov elevated the ‘objectivism’ of Marx in contrast to the subjective values of the *Narodniki*. He wrote, ‘the criterion of truth lies not in me, but in the relations which exist outside of me’. Thus, objectivity was possible in social theory. Plekhanov drew from Marx, and from the traditions of the English economists and German historicists, the fundamental principle that economic forces determine social development.

Plekhanov was active in the Second International (1889) and attended its Congresses in Zurich (1893), Amsterdam (1904) and Copenhagen (1910). Together with Lenin, Martov and Potresov, Plekhanov founded *Iskra* (The Spark) in 1900 – the first Russian Marxist newspaper. In 1903 he worked jointly with Lenin to draw up the programme adopted at the famous Second Congress of the Social Democratic Party, but it was shortly after this that Plekhanov broke with Lenin and the Bolsheviks and sided with the Mensheviks. During the Revolution of 1905.

Plekhanov advocated an ‘opportunist’ alliance with the liberals, while in 1914 he supported the war against Germany for the defence of Russia (in opposition to Lenin and the Bolshevik position). In that year he formed the *Yedinstvo* (Unity) group, which was designed to bring together the Mensheviks and the anti-Lenin Bolsheviks, but its influence was negligible.

After the Revolution of February 1917 Plekhanov returned to Russia, supporting the Provisional government and the continuation of the war. He denounced the Bolshevik coup of October 1917, and shortly afterwards fell ill with tuberculosis. Ostracized by Lenin and terrorized by the Cheka, Plekhanov’s wife took him to Finland, where he died on 30 May 1918.

Despite his differences with Lenin and the Bolsheviks after 1903, Plekhanov’s writings continued to be highly regarded and widely studied in the Soviet Union. During the 1920s his library and archives were gathered from a number of European centres and taken to Leningrad, where the Plekhanov Library was established, and his complete writings were published.

Selected Works

1883. Socialism and political struggle. In *Selected philosophical works*, vol. 1. Moscow: Progress Publishers, 1974.

1885a. Our differences. In *Selected philosophical works*, vol. 1. Moscow: Progress Publishers, 1974.

1885b. The development of the monistic view of history. In *Selected philosophical works*, vol. 1. Moscow: Progress Publishers, 1974.

1895. *Anarchism and socialism*. Trans. E. Marx Aveling. Chicago: C.H. Kerr.

1897. *Contributions to the history of materialism*. Trans. R. Fox as *Essays in the history of materialism*, London: Lane, 1934.

1923–7. *Sochineniya* [Works], 24 vols. Moscow.

1987. *Selected works*, 5 vols. London: Lawrence and Wishart.

Bibliography

Baron, S. 1963. *Plekhanov: The father of russian marxism*. London: Routledge & Kegan Paul.

Baron, S. 1996. *Plekhanov in russian history and soviet historiography.* Pittsburgh: University of Pittsburgh Press.

Chagin, B.A. 1963. *G.V. Plekhanov i ego rol’ v razvitii Marksistskoi Filosofii*. Moscow and Leningrad: Akademii Nauk.

Harding, N. 1977. *Lenin’s political thought, volume 1: Theory and practice of the democratic revolution*. New York: St. Martins.

Harding, N. 1981. *Lenin’s political thought, volume 2: Theory and practice of the socialist revolution*. New York: St. Martins.

Keep, J.L.H. 1963. *The rise of social democracy in Russia*. Oxford: Clarendon Press.

Kolakowski, L. 1978. *Main currents of marxism. Volume II: The golden age*. Oxford: Oxford University Press.

Pluralism in Economics

Sheila C. Dow

Abstract

This article explores the meaning of pluralism in economics and the arguments put forward in support of it. In particular, the distinction is drawn between methodological pluralism (support for variety in methodological approach) and a pluralist methodology (one which employs a variety of methods). Methodological pluralism usually takes the form of arguing that it is in the nature of knowledge about social systems that there will be variety of methodological approaches. But prescriptive arguments for a particular pluralist methodology may accompany the argument that this is the single best methodology (methodological monism).

Keywords

Critical realism; Methodological pluralism; Monism; Open systems; Paradigms; Pluralism in economics; Pluralist methodology; Positivism; Postmodernism; Rhetoric

JEL Classification

B4

Pluralism is the advocacy of plurality (Mäki 1997), or variety, and has been featured increasingly in discussions of economic methodology. Indeed, the International Confederation of Associations for Pluralism in Economics (ICAPE) is an umbrella organization for around 40 international economics organizations.

The term 'pluralism' was first used in modern methodological discourse by Bruce Caldwell (1982) in *Beyond Positivism*. Here he charted the growing dissent from a positivist methodology that had been put forward within a monist approach, that is, the advocacy of a single, best methodology. Positivism had prescribed progress in knowledge by means of empirical testing of propositions. Yet it had proved impossible to express all propositions in testable form, and difficult to derive definitive empirical tests for those propositions which were quantifiable (partly because of the so-called 'Duhem–Quine problem'). Caldwell concluded that, rather than searching for some other, elusive, monist methodological approach, economists should accept that a range of approaches could legitimately be sustained (though he later expressed the hope that there would some day be agreement on one best methodological approach: Caldwell 1989). We will use Caldwell's term 'methodological pluralism' for pluralism at this level of methodological approach.

The implication of methodological pluralism is that methodologists should study (critically) a range of methodologies rather than seek to identify one (universally) best methodology. Practising economists must, however, choose one methodological approach or another, but may nevertheless support methodological pluralism by accepting that, while they may have good reason for their choice of approach, they accept that its superiority cannot be demonstrably proven. This chosen methodology itself may (or may not) be pluralist, that is, employ a variety of methods. Similarly, there is scope for plurality at the theoretical level, whether or not there is methodological pluralism, or a pluralist methodology. Thus, for example, Colander (2000) draws the distinction between the theoretical pluralism of mainstream economics and its monism in terms of formalist method (see also Goodwin 2000). There has been considerable confusion in the literature between the meanings of pluralism at these different levels. No doubt this stems in part from the fact that many who support pluralism at one level also support it at another; but one does not necessarily entail either of the others.

Since Caldwell's initial proposal, a range of further arguments has been developed for methodological pluralism. Samuels (1997) has been a consistent exponent in practice of methodological pluralism, even before it had been explicitly identified (Caldwell 1997). His support arises from a critique of prescriptive epistemology, on the constructivist grounds that our knowledge of the economy is situated, and the economy itself is a social construction, so that there is no scope for a common methodological approach to knowledge. Samuels's argument is consistent with the postmodern critique that emphasizes the absence of independent facts by which to test theories, which follows from the subjective and fragmented perceptions of experience (a plurality of understandings). Samuels is explicitly prescriptive at the meta-methodological level – he positively advocates methodological pluralism (but that is the limit of his prescription). Postmodernists agree with Samuels that there is no basis for any form of limitation on plurality, but go further in arguing that there is no role for prescriptive methodology at all. It follows that methodological pluralism has no meaning for them, since it is a prescriptive position.

Weintraub (1989) and McCloskey (1983) draw the distinction between prescriptive, 'large M', Methodology and descriptive, 'small m', methodology. Thus the science studies and rhetoric approaches, respectively, see a role for the second in providing descriptive accounts of different methodological approaches. This takes the form of recognizing methodological plurality as a feature of the subject matter, while not advocating it. Although not prescriptively methodological pluralist themselves, Weintraub and McCloskey nevertheless support a weak form of pluralism, which takes the form of an ethical argument. If there is a plurality of methodological approaches to economics (that is, economics is not defined by a particular methodological approach), then our discourse should be structured in such a way as to take that on board. Thus there are injunctions to practising economists to respect the legitimacy of expressions of methodological difference, be polite and so forth (McCloskey 1996; Screpanti 1997).

While much of the support for pluralism has focused on knowledge limitations, others (including Caldwell 1997) have extended the argument to the nature of reality. Dating back to Keynes's (1921) *Treatise on Probability*, it has been argued that the organic nature of the social world means the absence of law-like behaviour. It is not just that the capacity for knowledge is limited, but that the basis for laws is not there to be found. Since the social world is believed to be open, it requires an open system of knowledge, one which allows both for change and variety in methodological approach (Chick and Dow 2005).

Keynes had argued further that reliable knowledge is derived from accessing a range of sources, given that there is so little scope for establishing demonstrably true knowledge. In the absence of certainty as to premises, classical logic is of limited use, so reason employs 'human logic'. While rationalism is inadequate as a basis for action, human logic rather involves drawing on evidence and theoretical knowledge, supplemented by conventional opinion and intuition. The weight of argument is greater, the greater the number of sources of knowledge which support the hypothesis. This is to be distinguished from the probability that the hypothesis is correct, which probability may rise or fall with new evidence. This is an argument for a pluralist methodology, that is, a methodology that employs a range of methods. Inevitably this means methods beyond mathematical formalism, since one of the main attractions of that method is that all arguments can be expressed commensurately and can therefore be collapsed into one formal argument, or model. But the reasoning can also be applied to the meta-methodological level, to support methodological pluralism. If a policymaker perceives support for a particular policy from a range of methodological approaches, weight is added to the view that the policy is a good one.

If no single methodological approach can be demonstrated to be universally the best one, then it is to be expected that there will be a range of methodological approaches, even without the added, Keynesian, argument that this is desirable. This range is bound to have some limits, given that science operates within loose communities, requiring shared understandings of reality (ontologies) and of meanings of terms, and shared views as to how to proceed to build up (fallible) knowledge. These communities can be understood in Kuhnian terms as paradigms (Kuhn 1962). This advocacy of variety of methodological approach limited according to the requirement for viable scientific communities is termed variously 'critical pluralism' (Caldwell

1997), 'principled relativism' (Davis 1999) or 'structured pluralism' (Dow 2004). Each of the range of methodological approaches involves a different set of views as to how best to build up knowledge. Some will be pluralist, advocating the use of a range of methods, but this does not necessarily follow from the application of methodological pluralism. There are trade-offs involved in whatever methodology is chosen, and the benefits of a single commensurate method may be judged to outweigh the epistemological costs. But those approaches which adopt a pluralist methodology will be distinguished by the selection of methods employed. This in turn follows from ontology, the understanding of the nature of the subject matter. An understanding of economic relations in terms of class implies one set of methods, in terms of competitive markets another, and of individual entrepreneurial creativity another, for example.

The focus on the ontological level for the case for pluralism owes much to critical realism (Lawson 1997, 2003). However, critical realists themselves take a particular position on pluralism. The case is made that the real world is open, requiring an open-system epistemology. The open-system nature of the real world means that knowledge is situated and contestable, and therefore there is likely to be a range of methodologies; to that extent critical realism is methodologically pluralist. But, since there is one external real world, there is only one ontology, and one open-system meta-methodology. Different methodologies simply reflect different 'commitments' with respect to that common ontology; beyond that no judgement is to be expressed on the content of these methodologies. This approach differs from the structured pluralist position outlined above, whereby different methodological approaches follow ultimately from the plurality of ontological understandings (with respect to a common external real world). However, critical realists also support the idea that methodologies should be pluralist, adopting a range of methods suited to the chosen focus of analysis.

Finally, the argument that plurality of methodology is not only inevitable but desirable is supported by means of a biological metaphor (Hodgson 1997). The view that the real social world is open involves the view that it undergoes structural change. Even if it were generally agreed that a particular methodological approach is best suited to the current economic structure, it is not unlikely that this approach would not be capable of addressing change in that structure. In the biological world, a dominant strain of a species may not be able to survive an environmental shock. Unless there is a range of alternative strains, including one better suited to the new environment, the species will die out. Since the nature of environmental shocks in general cannot be predicted, it is necessary for the survival of the species for there to be a range of alternative strains available at any time. The same argument can be made for different methodological approaches to economics.

See Also

- ▶ Keynes, John Maynard (1883–1946)
- ▶ Methodology of Economics
- ▶ Paradigms
- ▶ Scientific Realism and Ontology
- ▶ Theory Appraisal

Bibliography

Caldwell, B.J. 1982. *Beyond positivism*. London: Allen & Unwin.

Caldwell, B.J. 1989. Post-Keynesian methodology: An assessment. *Review of Political Economy* 1: 43–64.

Caldwell, B.J. 1997. Comment. In Salanti and Screpanti (1997).

Chick, V., and S.C. Dow. 2005. The meaning of open systems. *Journal of Economic Methodology* 12: 363–381.

Colander, D. 2000. New millennium economics: How did it get this way, and what way is it? *Journal of Economic Perspectives* 14(1): 121–132.

Davis, J.B. 1999. Postmodernism and identity conditions for discourses. In *What do economists know? New economics of knowledge*, ed. R.F. Garnett Jr.. London: Routledge.

Dow, S.C. 2004. Structured pluralism. *Journal of Economic Methodology* 11: 275–290.

Goodwin, C. 2000. Comment: It's the homogeneity, stupid! *Journal of the History of Economic Thought* 22: 179–184.

Hodgson, G.M. 1997. Metaphor and pluralism in economics: Mechanics and biology. *Pluralism in economics*. In Salanto and Screpanti (1997).

Keynes, J.M. 1921. *A treatise on probability. Collected writings*, vol. 3. London: Macmillan for the Royal Economic Society, 1973.

Kuhn, T.S. 1962. *The structure of scientific revolutions*. Chicago: University of Chicago Press.

Lawson, T. 1997. *Economics and reality*. London: Routledge.

Lawson, T. 2003. *Reorienting economics*. London: Routledge.

Mäki, U. 1997. The one world and the many theories. *Pluralism in Economics*. In Salanti and Screpanti (1997).

McCloskey, D.N. 1983. The rhetoric of economics. *Journal of Economic Literature* 21: 434–461.

McCloskey, D.N. 1996. *The vices of economists: The virtues of the Bourgeoisie*. Amsterdam: Amsterdam University Press.

Salanti, A., and E. Screpanti (eds.). 1997. *Pluralism in economics*. Cheltenham: Edward Elgar.

Samuels, W. J. 1997. The case for methodological pluralism. In Salanti and Screpanti (1997).

Screpanti, E. 1997. Afterword: Can methodological pluralism be a methodological canon? In Salanti and Screpanti (1997).

Weintraub, E.R. 1989. Methodology doesn't matter, but the history of thought might. *Scandinavian Journal of Economics* 91: 477–493.

Plutology (Gr. πλοῦοσ, Wealth)

A. W. Flux

This term was used by Courcelle-Seneuil to describe that part of his treatise on political economy which dealt with what is described by some more modern writers as ‘pure theory’; that scientific study of the results of the action of economic motives on men and societies to which the terms ‘economics’ and ‘economic science’ have been applied in the effort to escape the confusions which arose from embracing under the general title) ‘political economy’, both these more abstract investigations and the application of the knowledge thus gained, with that derived from other sources, to problems of practical statemanship. To this second part of the subject the eminent French economist applied the term *Ergonomy*. The Australian W.E. Hearn adopted the title for his work, *Plutology, or the Theory of the Efforts to satisfy Human Wants*.

See Also

- ▶ Donisthorpe, Wordsworth (1847–1914)
- ▶ Hearn, William Edward (1826–1888)

Bibliography

Courcelle-Seneuil, J.G. 1858. *Traité théorique et pratiqueold" d'économie politique*. Paris: Guillaumin.

Hearn, W.E. 1864. *Plutology, or the theory of the efforts to satisfy human wants*. London: Macmillan.

Polak, Jacques Jacobus (Born 1914)

S. C. Tsiang

Born in Rotterdam, Polak was educated at the University of Amsterdam, where he obtained an MA (Econ.) degree in 1936, and a PhD (Econ.) in 1937. His early research work as an economist with the League of Nations was concerned with the study of business cycles and he assisted Jan Tinbergen in the latter's work on business cycles under the sponsorship of the League of Nations. Later they co-authored Polak and Tinbergen (1950).

After World War II Polak moved to the IMF, where he worked first as the chief of the statistics division, and then as the deputy director, and eventually became the director of the research department. His research interests then turned towards the monetary analysis of income determination and the balance of payments. His work in this area to some extent anticipated later developments in the monetary theory of the balance of payments by the Chicago School, although he himself might have been influenced by the earlier works of his colleagues at the IMF, for example by S.C. Tsiang's studies of the balance of payments improvement of Denmark in 1951 and the exchange reform of Peru (1950–54).

As a prominent IMF official, he was deeply involved in the mid-1960s in the design of the Special Drawing Rights (SDR), which were an application of the original Keynesian suggestion of 'Bancor' and which greatly expanded the lending capacity of the Fund.

In 1981 Polak became Executive Director for Cyprus, Israel, the Netherlands, Romania and Yugoslavia at the International Monetary Fund (IMF), Washington, DC.

Selected Works

1939. The international propagation of business cycles. *Review of Economic Studies* 6: 79–99.

1943. Balance of payments problems of countries reconstructing with the help of foreign loans. *Quarterly Journal of Economics* 57. Reprinted in American Economic Association, *Readings in the theory of international trade*. Philadelphia: Blakiston, 1949.

1950. (With J. Tinbergen.) *The dynamics of business cycles*. Chicago: University of Chicago Press.

1954. *An international economic system*. Chicago: University of Chicago Press.

1957. Monetary analysis of income formation and payments problems. *IMF Staff Papers* 6: 1–50.

1960. (With Lorette Boissonnealt.) Monetary analysis of income and imports and its statistical application. *IMF Staff Papers* 7: 349–451.

1970. Money: National and international. In *International reserves: Needs and availability*, ed. IMF, 510–520.

1977. (With R.A. Mundell, eds) *The new international monetary system*. New York: Columbia University Press.

1984. The rôle of the International Monetary Fund. In *Problems of the international monetary system, forty years after Bretton Woods*. Boston: Federal Reserve Bank of Boston.

Bibliography

Tsiang, S.C. 1953. The 1951 improvement in the Danish balance of payments. *IMF Staff Papers* 3.

Tsiang, S.C. 1957. An experiment with a flexible exchange rate system: The case of Peru, 1950–1954. *IMF Staff Papers* 5.

Polanyi, Karl (1886–1964)

George Dalton

Keywords

Machine technology; Managed capitalism; Polanyi, K.; Pre-industrial economies

JEL Classifications

B31

Polanyi was born in Vienna in 1886 and grew up in Budapest, where he studied law and philosophy. He served as an officer in the First World War, after which he turned to economic journalism as foreign editor of Vienna's *Osterreichisiche Volkswirt* throughout the 1920s. He emigrated to England in 1933, where he worked in adult education, as a lecturer on world affairs for the Workers' Educational Association and for the Extramural Delegacies of the Universities of Oxford and London. He became intensely interested in the origins of the British Industrial Revolution and the enormity of its economic and social consequences, the subject of his book, *The Great Transformation* (1944), written while he was a resident scholar at Bennington College in Vermont between 1940 and 1943.

John Maurice Clark was sufficiently impressed by the book to invite Polanyi to Columbia University as a visiting professor of economic history in 1947 (when Polanyi was already 61). Polanyi remained at Columbia until his retirement in 1953. He continued doing research until his death in 1964 at his home in a suburb of Toronto.

The Great Transformation remains in print 45 years after its publication. It argues a triple thesis: (i) that in Great Britain and Western Europe, the coming of machine technology to mercantilistic national economies that contained governmentally regulated markets induced enormous growth in all input and output markets and the removal of governmental controls from some of them – what Polanyi calls an attempt to create a 'self-regulating' market system; (ii) that nationally integrated market systems in which labour, land, and money as well as produced goods were transacted as market commodities were historically unique (that such full-blooded capitalism dominated by market transactions for factor inputs as well as produced outputs was a new kind of economic system markedly different from any that preceded it); (iii) although machine technology producing within a market system was enormously productive – an 'unbound Prometheus' in the vivid phrase used by David Landes – its destructive consequences (that is, sporadic unemployment, the business cycle, large inequalities in income and wealth) culminating in the Great Depression of the 1930s, forced governments from the early 19th century onwards to initiate market controls, monetary and fiscal policy to mitigate its destructive consequences, what we now call 'managed' and 'welfare state capitalism'.

Polanyi's second big book, *Trade and Market in the Early Empires* (1957), which also remained in print after 30 years and which has also been translated widely, created a theory of pre-industrial, non-market economies of interest to economic archaeologists, economic anthropologists, and those economic historians who study early, pre-industrial economies throughout the world. Polanyi invented a conceptual vocabulary to specify the core attributes of such early and primitive economies much of which is employed today in standard fashion: 'reciprocity', 'redistribution', 'special-purpose money', 'port of trade', 'politically administered trade', 'economy embedded in society'. This part of Polanyi's work is widely thought to illuminate the nature of early money, early foreign trade, and the economic organization of early kingdom-states.

Polanyi's continuing significance is reflected in the Karl Polanyi Institute of Political Economy, founded at Concordia University, Canada, in 1987, which in addition to scholarly activity that is motivated by Polanyi's thought, maintains an archive of his works.

Selected Works

1944. *The great transformation.* New York: Rinehart.

1957. *Trade and market in the early empires*, ed. C.-M. Arensberg, and H.W. Pearson. Glencoe: Free Press.

1966. *Dahomey and the slave trade*. Seattle: University of Washington Press.

1971. *Primitive, archaic, and modern economies: Essays of Karl Polanyi*, ed. G. Dalton. Boston: Beacon Press.

1977. *The livelihood of man*. New York: Academic Press.

1996. *Uncollected works of Karl Polanyi*. New York: St. Martins.

Bibliography

Adelman, I., and C.T. Morris. 1978. Patterns of market expansion in the nineteenth century: A quantitative study. In *Research in economic anthropology*, ed. G. Dalton, vol. 1. Greenwich: JAI Press.

Dalton, G., and J. Köcke. 1983. The work of the Polanyi group: Past, present, and future. In *Economic anthropology*, ed. S. Ortiz. New York: University Press of America.

Polanyi-Levitt, K., ed. 1990. *The life and work of Karl Polanyi*. Montreal: Black Rose Books.

Stanfield, J. 1986. *The economic thought of Karl Polanyi: Lives and livelihood*. London: Palgrave MacMillan.

Valensi, L., et al. 1981. Economic anthropology and history: The work of Karl Polanyi. In *Research in economic anthropology*, ed. G. Dalton, vol. 4. Greenwich: JAI Press.

Polarization

Gordon Anderson

Abstract

Polarization means the tendency of economic agents to form different groups and acquire identities that enhance differences from other groups. It is both cause and consequence of much economic behaviour. It has been employed, for example, in describing the diminution of the middle class in wage, income and wealth distributions, in studying growth and convergence issues, and in examining the plight of the poor. Although polarization is closely associated with trends in inequality, increased polarization can correspond to an increase, a reduction or no change in inequality.

Keywords

Akerlof, G,; Alienation; Clubs; Endogeneous growth; Gini coefficient; Identity; Inequality; Polarization

JEL Classifications

D3; I3

Polarization, or the tendency of economic agents to collect into different groups and to feel increasingly different from members of other groups, is both cause and consequence of much economic behaviour. The essence of informal and formal club formation, polarization is born of an increasing sense of identity within the group members and an increasing sense of distance from members of other groups. The terminology is gaining increasing currency in economics. Akerlof (1997); Anderson (2004a, b); Beach and Slotsve (1996); Beach et al. (1998); Bossert et al. (2004); Corak (2004); D'Ambrosio and Wolff (2001); Dinardo and Lemieux (1997); Foster and Wolfson (1992); Jenkins (1996); Jones (1997); Keefer and Knack (2002); Levy and Murnane (1992); Quah (1997) and Wolfson (1994, 1997) constitute an extensive but not exhaustive list of its use. In the list will be seen applications in describing the diminution of the middle class in wage, income and wealth distributions, in studying growth and convergence issues, and in examining the plight of the poor; it has also been used in the study of inter-generational income relationships and in discriminating between competing matching models of marriage partners. These literatures broadly interpret polarization as the disappearance of mass at the centre of an empirical distribution of a characteristic, or the increasing distance between, and intensity of, multiple points of modality of the distribution as it evolves through time. It is inherently a dynamic process involving the comparison of the anatomy of states at different points in time, essentially examining how the shape of the distribution of a characteristic (or a collection of characteristics) has evolved during the process. Thus, the objective is to detect trends in shapes of distributions over time that reflect the polarization or de-polarization (sometimes referred to as 'convergence') of that group of agents. Although polarization is closely associated with trends in inequality, it is distinguishable and quite different from changes in inequality in that increased polarization can induce an increase, a reduction or no change in inequality.

The concept need not be confined to the study of changes within a population's distribution of a particular characteristic, but can be used in assessing the relative movements of two or more distributions as they evolve (for example, polarization between ethnic groups, genders, and nations). In this context polarization takes the form of distributions becoming 'less alike' in a particular fashion; as such it involves comparison of complete distributions, not just their location or scale characteristic. While the identification of polarization within and between populations presents quite distinct empirical challenges, polarizations have many common features which can be exploited in understanding the nature of the phenomenon. Indeed, it is convenient to contemplate within-distribution polarization phenomena as the consequence of that population distribution being a mixture of sub-population distributions which are themselves polarizing (this notion is at the heart of the initial formalization of a Polarization index in Esteban and Ray 1994, and Duclos et al. 2004). Such a construction will highlight why within distribution polarization is sometimes hard to detect in the absence of sub-population information.

Polarization: An Axiomatic Foundation

Indices of polarization were formulated in Esteban and Ray (1994) and Duclos et al. (2004) (see also Wang and Tsui 2000) by positing a collection of axioms whose consequences should be reflected in a Polarization measure. The axioms are founded upon a so-called Identification–Alienation nexus wherein notions of polarization are fostered jointly by an agent's sense of increasing within-group identity and between-group distance or alienation. The four axioms may be loosely summarized as follows:

Axiom 1 A mean preserving reduction in the spread of a distribution cannot increase polarization.

Axiom 2 Mean preserving reductions in the spread of sub-distributions at the extremes of a density cannot reduce polarization.

Axiom 3 Separation of two sub-densities towards the extremes of the distributions range must increase polarization.

Axiom 4 Polarization measures should be population-size invariant.

The polarization index developed for discrete distributions as a consequence of these axioms (Esteban and Ray 1994) may be written as:

$$P_\alpha = K\sum_{i=1}^{n}\sum_{j=1}^{n}\left|x_i - x_j\right|\pi_i^{1+\alpha}\pi_j \qquad (1)$$

Here K is a normalizing constant, π_i is the sample weight of the i'th observation and where $\alpha \geq 0$ is a polarization sensitivity factor chosen by the investigator. It may readily be seen that $\alpha = 0$ yields a sample weighted Gini coefficient.

The continuous distribution analogue (Duclos et al. 2004) may be written as:

$$P_\alpha(F) = \int_y f(y)^\alpha \int_x |y - x| dF(x)dF(y) \qquad (2)$$

Again, α is the polarization sensitivity factor which in this case is confined to [0.25,1].

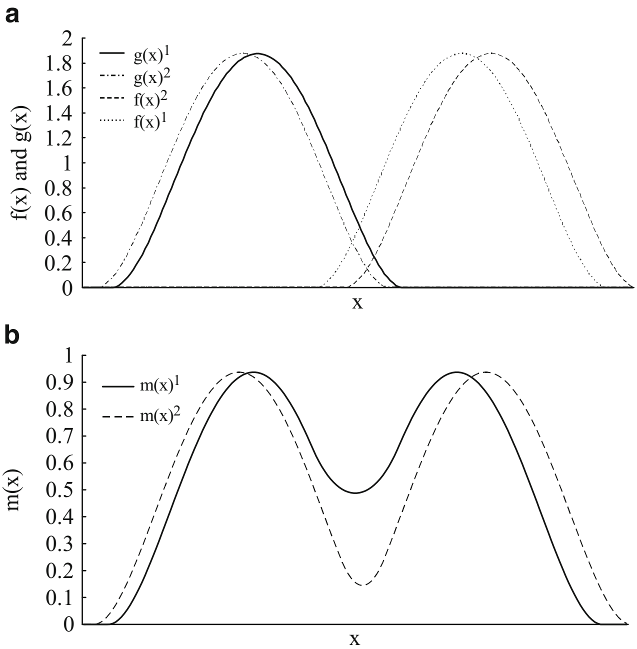
The Anatomy of Polarized States

For expositional simplicity in exploring the anatomy of polarized states, a population is represented by the equi-probable mixture of two sub-distributions (in reality more than two subgroups are possible), representing two subgroups or clubs which make up the population. The initial sub-distributions, which are identical except for having different means, are subjected to various transformations which are characterized in Figs. 1a, b, 2a, b, 3a, b and 4a, b.

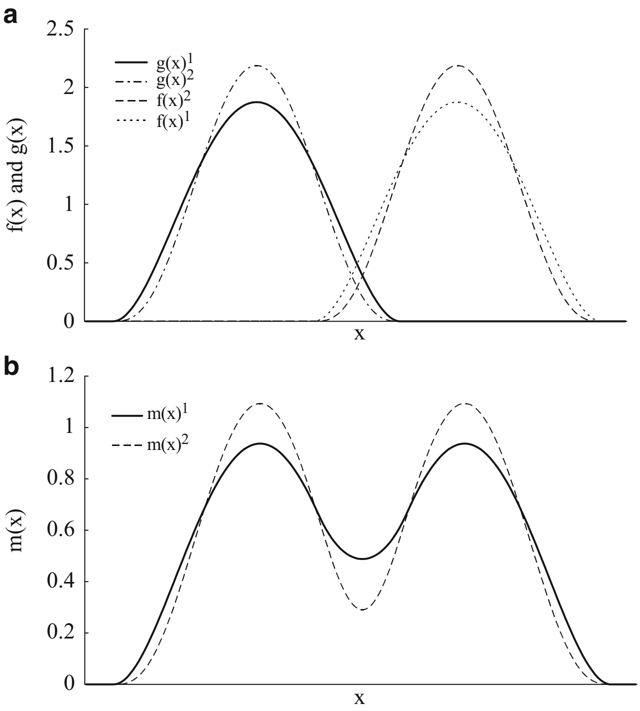
Following the spirit of Axiom 3, the simplest form of polarization occurs when the sub-populations exhibit divergence in their means. Here there is no increase in within-group identity but there is an increase in the distance between members of different groups (alienation). Figure 1a exemplifies this situation in terms of the sub-populations and its consequence for the mixture is illustrated in Fig. 1b. As may be seen, the overlap of the two sub-distributions diminishes, and the centre of the mixture becomes hollowed out. Note that when the means are relatively close together there is no hollowing out but simply a flattening of the unimodal peak of the mixture distribution, implying that polarized or polarizing states need not be characterized by the existence or emergence of bimodality. (For example, for mixtures of equal-variance normal distributions, bimodality will not emerge under any mixing scheme until the difference in means exceeds $4.5^{0.5}$ standard deviations). Thus, bump-hunting techniques available in the statistics literature (Good and Gaskins 1980; Hartigan and Hartigan 1985) which seek out inflections in the probability density function will not necessarily be useful in the analysis of polarization.

Figure 2a, b illustrate another form of polarization when sub-population means remain constant but their variances diminish. This is much in the spirit of Axiom 2 and characterizes a situation of increased identification within the groups without an increased sense of alienation between them. Again, the overlap of the sub-populations diminishes and the centre of the mixture becomes hollowed out, but in this case the anatomical change is not unequivocal. Finally, when both locations and spreads remain constant but the lower distribution skews left and the upper distribution skews right, the overlap again diminishes

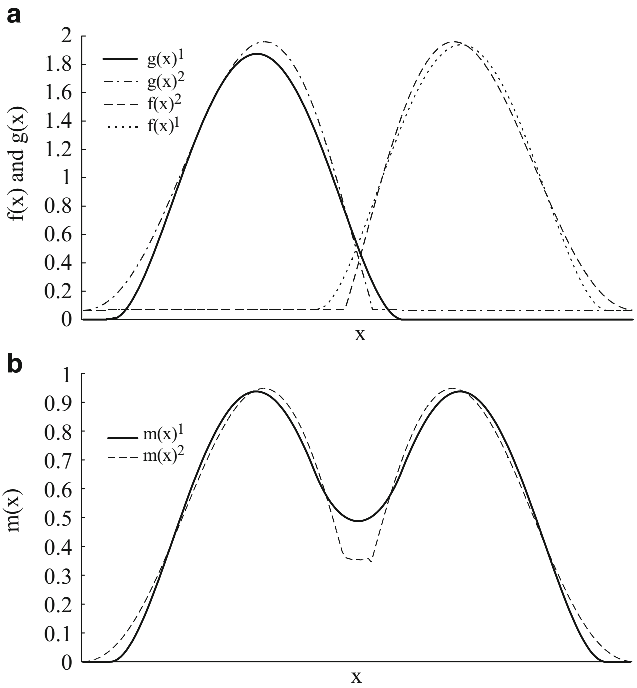
Polarization, Fig. 1
(**a**) Divergence in means between population polarization.
(**b**) Divergence in means within population polarization



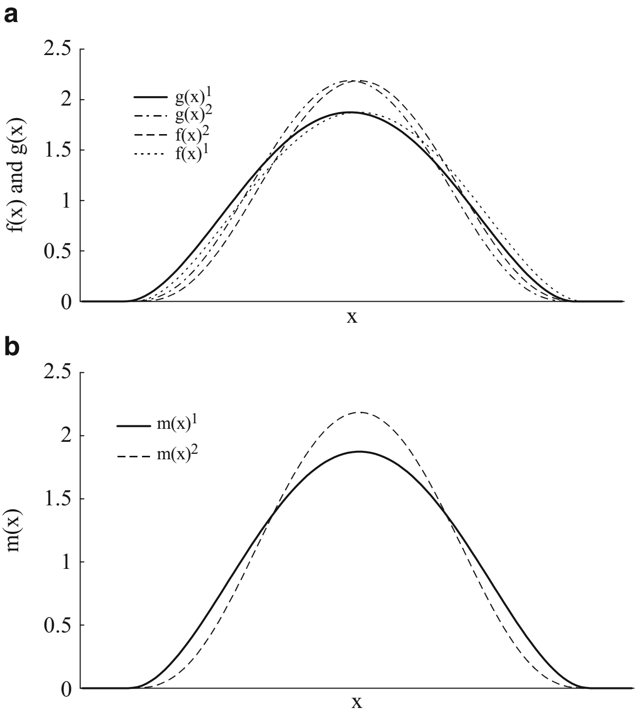
Polarization, Fig. 2
(**a**) Increased concentration between population polarization. (**b**) Increased concentration within population polarization



Polarization, Fig. 3
(**a**) Opposite skewness between population polarization. (**b**) Opposite skewness within population polarization



Polarization, Fig. 4
(**a**) Increased concentration between population polarization close means.
(**b**) Increased concentration within population polarization close means



and the centre of the mixture distribution again hollows out, as Fig. 3a, b illustrate.

One thing these examples demonstrate is the potential for changes in the overlap measure to provide a general test or indicator of polarization. But this is not without qualification. Figure 4a, b return us to the increased identification case and demonstrate that, if the subgroup means happen to be close together, the extent of overlap can increase and the mass at the centre of the mixture distribution is enlarged with increased polarization. This highlights what is in effect a potential statistical identification problem associated with polarization when only the mixture of sub-distributions is observed. The tenuous link between polarization and inequality is also illustrated in this example. If we consider the mixture $f(x)$ to be an equally weighted mixture of normal distributions $N\left(\mu_\mathrm{i}, \sigma_\mathrm{i}^2\right), \mathrm{i} = 1, 2$, then the variance of x (for our purposes a measure of inequality) which will be $0.5\left(\sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2\right)$ can be seen to either increase, decrease or remain unchanged with an increase in polarization interpreted as any combination in reductions of sub-population variances and increases in the difference of sub-population means.

Alternative Between-Group Polarization Measures

Distributional Overlap

The anatomy analysis suggests that one technique for assessing polarization between two groups is to evaluate how much they have in common. Such a measure corresponds to non-alienation, and its negative (or some negative function of it) corresponds to a degree of alienation. Anderson (2004a, b) proposes an overlap measure as an index of convergence and a function of its negative as a measure of alienation. The extent to which two distributions $f(x)$ and $g(x)$ overlap is given by:

$$OV = \int_{-\infty}^{\infty} \min(f(x), g(x))dx \qquad (3)$$

Clearly it is a number between 0 and 1 with 0 corresponding to no overlap and 1 to the perfect matching of the two distributions. It follows that $1-\ OV$ is a measure of the extent to which the distributions do not match or are alienated. When $f(x)$ and $f\,(y)$ are specified to the extent that all of their parameters can be estimated and the intersection points of $f(x)$ and $g(x)$ calculated, OV can be estimated parametrically (see Anderson and Ge 2004). When $f(\cdot)$ and $g(\cdot)$ are unknown, given independent samples from $f(\cdot)$ (represented by x) and $g(\cdot)$ (represented by y) of sizes n_x and n_y respectively, its empirical counterpart may be implemented by choosing $K + 1$ mutually exclusive and exhaustive partitions of the range of x whose upper bound is defined by x^k, $k = 1, \ldots, K + 1$ and calculating

$$OV^e = \sum_{i=1}^{k+1} \min\left(\frac{\sum_{j=1}^{n_x} I\left(x_j, x^i\right)}{n_x}, \frac{\sum_{j=1}^{n_y} I\left(y_j, x^i\right)}{n_y}\right) \qquad (4)$$

where $I(z,w^\mathrm{i})$ is an indicator function equal to 1 when z is in the interval $(w^{i\,-\,1}, w^\mathrm{i})$ and 0 otherwise. The statistical properties of such an estimator are discussed in Anderson (2005). The nonparametric measure is prone to two sources of bias: the first, due to the intersection points of the underlying distributions not coinciding with partition points, is actually not that large provided k is not small and the partition points are chosen judiciously; the second, due to the estimator being implicitly a conditional estimator, can be large when the measure is either close to 0 or 1. However, these biases do not appear to impede its use in calibrating changes in overlap. The main problem with this particular instrument arises when distributions do not actually overlap; for that purpose the following measures may prove useful.

Gini-Based Between-Group Polarization Measures

Starting with the classic Gini inequality coefficient which, with x_i being the income of the i'th agent for agents $i = 1,\ldots,n$ and where for convenience and without loss of generality incomes are arranged in ascending rank order, may be written as:

$$Gini = \frac{1}{2n^2\mu}\sum_{i=1}^{n}\sum_{j=1}^{n}\left|x_i - x_j\right| \qquad (5)$$

where μ is the mean of the *x*'s. Suppose the rich and poor groups are defined by a poverty cut-off somewhere between x_p and x_{p+1} where $1 < p < n$ (what Yitzhaki 1994 refers to as perfect stratification of groups, that is, no overlapping): then Gini may be thought of as the sum of the average mean normalized differences between agents in the poor group, between agents in the rich group, and between poor- and rich-group agents. In measuring alienation it is only the last group of comparisons that are relevant, that is, the average normalized difference between the rich-group and poor-group agents. In this case the new 'PGini' index could be written as:

$$PGini = \frac{1}{p(n-p)\mu}\sum_{j=p+1}^{n}\sum_{i=1}^{p}\left(x_j - x_i\right) \qquad (6)$$

Clearly this is still a number greater than 0 (but it is no longer guaranteed to be less than 1), which reflects the mean normalized average distance between the poor group and the rich group, and as such it is easy to show that it is the overall mean normalized difference between the subgroup means. Indeed, the formulae can be generalized to general group differences where there is not perfect segmentation, that is, where the subgroups overlap.

Observe that the same index would be arrived at if one were to work with $x_i - z$ and $x_j - z$, the corresponding distances from the poverty line *z* which facilitates a link to the well-known FGT family of poverty and welfare indices introduced by Foster et al. (1984), as follows. The formal representation of this family is given by:

$$POV_\theta(x,z) = \int_0^z \left(\frac{z-x}{z}\right)^\theta dF(x) \qquad (7)$$

where *F*(*x*) is the cumulative density function (with p.d.f. *f*(*x*)) describing the population of incomes, *z* is the maximum of the poor, and $\theta(\geq 0)$ is a parameter defining the nature of the poverty index and corresponds to a measure of poverty aversion. As a consequence POV_0 corresponds to the proportion of people in the poverty group,POV_1 is a normalized measure of the intensity of relative deprivation and so on. POV_i/POV_0 may be construed as the expected value of a weighted function of the normalized income deficiency where the weights are the $(i-1)$'th power of the normalized income deficiency itself. Thus increasing *i* increases the weights attached to those furthest from the poverty line. Interestingly, as *i* becomes very large the index becomes a Rawlsian measure in focusing almost entirely on the poorest agent. All of these measures obey the focus axiom, which holds that poverty measures should depend only upon the incomes of the poor. As such they are not in any sense related to the status of the rich.

Along similar lines *RIC*(*x,z*), an index of weighted relative distances of incomes above the poverty line, may be contemplated whose theoretical representation is of the form:

$$RIC_\theta(x,z) = \int_z^{xmax} \left(\frac{x-z}{z}\right)^\theta dF(x) \qquad (8)$$

In this case *x* max corresponds to the maximum possible income. Here RIC_0 corresponds to the proportion of the population above the poverty line, RIC_1 is a normalized measure of relative well-being of the non-poor, RIC_2 is a measure of the intensity of the relative well-being of those above the poverty line, and so on. In this case as α becomes very large the index becomes almost entirely focused on the richest person, RIC_1/RIC_0 corresponds to the expected normalized income surplus over the maximum poverty income, and so on. For all $\theta > 0$ all of these indices are essentially measuring relative weighted distances from the poverty line, and it is in this sense that they are considered relative measures. However, both *RIC* and *POV* are completely uninformed with respect to the distribution of incomes in the other group, which accords with the focus axiom mentioned above. For the purposes of reflecting the notion of alienation between the poor and non-poor groups, this axiom needs to be violated. Indeed the population analogue of PGinican be shown to be a specific member ($\theta = 1$) of a general class of polarization measures defined by

$$POL(z,\theta) = \left(\frac{z}{\mu}\right)^{\theta}\left(\frac{RIC_\theta}{RIC_0} - \frac{POV_\theta}{POV_0}\right) \tag{9}$$

where $\theta \geq 1$.

Tests for Polarization

Given the distribution of the above indices, tests for increases or decreases in polarization in terms of movements in the indices can be readily established. But, although indices provide complete orderings, much like the Gini coefficient with which they are associated, they can be ambiguous. Direct tests of the anatomy of polarization based upon degrees of separation or stochastic dominance between density functions can provide an unambiguous (though not complete) orderings of the states of polarization. These tests can be developed by employing combinations of stochastic-dominance conditions, tests for which have been proposed by Anderson (1996, 2004a), Davidson and Duclos (2000), Barrett and Donald (2003), Linton et al. (2002), and McFadden (1989). The conditions can be used in combination to compare the right separation of the upper distribution with the left separation of the lower distribution and thus establish a statistical criterion for polarization both within and between distributions. Anderson (2004b) provides a taxonomy of such tests.

An Alternative Approach: The Growth and Convergence Literature

The endogenous growth literature has for a long while been concerned about issues of polarization specifically in the form of convergence or de-polarization. Early attempts at identifying the phenomenon via panel data regression techniques (see, for example, Barro 1998) ran into difficulties in interpretation (Bernard and Durlauf 1996). The phenomenon has been studied via the use of probability transition matrices implicit in the Markov chain methods employed by Quah (1997) (see also Durlauf and Quah 1999). (These techniques have also been applied to the problem of intergenerational Income relationships; see Corak 2004, and city sizes, Dobkins and Ioannides 2000, Anderson and Ge 2004). If we let *f*(*y*) be the distribution of income *y* in some future period, and *f*(*x*) the distribution of income *x* in the present period, the issue to be addressed is the relationship between the two distributions, that is, the extent to which, and manner in which, *f*(*y*) and *f*(*x*) are related. If for the moment we think of *x* and *y* as having the joint distribution *f*(*y*,*x*) so that *f*(*y*) and *f*(*x*) are the respective marginal distributions, at one extreme there is a sense of no relationship – that is to say, when *x* and *y* are independent $f(y, x) = f(y)f(x)$ – at the other there is the completely deterministic environment whereby $y = a + bx$ and the joint distribution is degenerate. If *y* and *x* are partitioned into *k* mutually exclusive and exhaustive regions where *p*(*y*) and *p*(*x*) are respectively the vectors of marginal probabilities of falling into those regions, interest centres on the elements of the square matrix *T* defined by $p(y) = T(y, x)p(x)$, the matrix in the square brackets in the following equation. *T* is of course the matrix of conditional probabilities formed by the product of the two square matrices in the equation, so that:

$$\begin{pmatrix} p_1(y) \\ p_2(y) \\ . \\ p_k(y) \end{pmatrix} = \left[\begin{pmatrix} p_{11}(y,x) & p_{12}(y,x) & . & p_{1k}(y,x) \\ p_{21}(y,x) & p_{22}(y,x) & . & p_{2k}(y,x) \\ . & . & . & . \\ p_{k1}(y,x) & p_{k2}(y,x) & . & p_{kk}(y,x) \end{pmatrix}\begin{pmatrix} p_1(x) & 0 & & 0 \\ 0 & p_2(x) & & 0 \\ . & . & . & . \\ 0 & 0 & . & p_k(x) \end{pmatrix}\right] \times \begin{pmatrix} p_1(x) \\ p_2(x) \\ . \\ p_k(x) \end{pmatrix} \tag{10}$$

which is a matrix of conditional probabilities – that is, $T = \|p_{ij}(y, x)/p_j(x)\|i$, $j = 1, \ldots,$ k–familiar in the convergence literature and made popular by Quah (1997). As such its properties are well known, as are the techniques for its estimation. The i'th column of T is a conditional probability density function describing the distribution or reallocation over states of the i'th element of $p(x)$ the initial income distribution, to the elements of $p(y)$, the resultant income distribution, after one period. If this process is thought to be time invariant, then letting p_s be the vector of $p_i(x)$'s s periods ahead, $p^s = T^s p$ corresponds to the s period ahead distribution, and the solution to $p^\infty = Tp^\infty$ (if it exists) is what is known as the long-run ergodic mass function. By interpreting these ergodic distributions as 'characterizations of tendencies', one can infer a tendency towards polarization if they display multiple peaks. Polarization can be examined in two ways in this context. When the diagonal elements of T are large relative to the off-diagonal elements, the system is said to exhibit persistence; if the diagonal is particularly large in the high and low ends it indicates a tendency towards polarization. Alternatively, one could compare p^∞, the long-run distribution, with $p(x)$, the initial distribution. If the former exhibits multiple peaks whereas the latter does not, a polarizing tendency may be inferred. One difficulty here is that no theory of inference has as yet been outlined for examining the 'multiple peakedness' of these ergodic functions.

Multivariate Polarization

When agents are characterized in terms of more than one characteristic, their polarization or otherwise will be reflected in more than one dimension. The empirical problem is then altogether much more challenging; the extension of the analysis to a multivariate measures can be somewhat problematic. Multivariate Gini coefficients have been developed (see Anderson 2004c, and Koshevoy and Mosler 1997) but adapting them to the current context is complex; it requires defining a poverty cut-off for each characteristic (or some poverty boundary in multidimensional space), but even then extending the analogy to multivariate measures of FGT indices is not possible. One simple approach is to take a weighted geometric mean of the various Gini coefficients in each dimension; but then the weights have to be determined in an inevitably arbitrary fashion.

On the other hand, extension of the overlap measure OV to a multivariate overlap measure MOV is very straightforward, since MOV is of the form:

$$MOV = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\min(f(x,y\ldots,z),\; g(x,y\ldots,z))dxdy\ldots dz \tag{11}$$

In the corresponding empirical measure MOV^e, given suitable partitions in each dimension, the indicator function would simply be modified to a multivariate version accordingly (Anderson 2005, provides an example) and $1 - MOV$ would provide an appropriate polarization measure.

See Also

- Alienation
- Convergence
- Gini Ratio
- Inequality Between Nations
- Wage Inequality, Changes in

Bibliography

Akerlof, G. 1997. Social distance and social decision. *Econometrica* 65: 1005–1027.

Anderson, G. 1996. Nonparametric tests for stochastic dominance in income distributions. *Econometrica* 64: 1183–1193.

Anderson, G. 2004a. Making inferences about the polarization, welfare and poverty of nations: A study of 101 countries 1970–1995. *Journal of Applied Econometrics* 19: 537–550.

Anderson, G. 2004b. Toward an empirical analysis of polarization. *Journal of Econometrics* 122: 1–26.

Anderson, G. 2004c. *The empirical assessment of multidimensional inequality: Sample weighted multivariate generalizations of the Gini Coefficient and Kolmogorov-Smirnov two sample tests for stochastic dominance*. Mimeo: Department of Economics, University of Toronto.

Anderson, G. 2005. *Indices and tests for alienation based upon Gini type and distributional overlap measures*. Mimeo: Department of Economics, University of Toronto.

Anderson, G., and Y. Ge. 2004. *A new approach to convergence, city types and the 'true' convergence of post-reform Chinese urban income distributions*. Mimeo: Department of Economics, University of Toronto.

Barrett, G., and S. Donald. 2003. Consistent tests for stochastic dominance. *Econometrica* 71: 71–104.

Barro, R. 1998. *Determinants of economic growth: A cross country empirical study*. Cambridge, MA: MIT Press.

Beach, C., R. Chaykowski, and G. Slotsve. 1998. Inequality and polarization of male earnings in the US 1968–1992. *North American Journal of Economics and Finance* 8: 135–151.

Beach, C., and G. Slotsve. 1996. *Are we becoming two societies? income, polarization and the middle class in Canada*. Toronto: C.D. Howe Institute.

Bernard, A., and S. Durlauf. 1996. Interpreting tests of the convergence hypothesis. *Journal of Econometrics* 71: 161–174.

Bossert, W., C. D'Ambrosio, and V. Peragrine. 2004. *Deprivation and social exclusion. Paper presented at the International Association for Research in Income and Wealth*. Ireland, August: Cork.

Corak, M. 2004. *Generational income mobility in north America and Europe*. Cambridge: Cambridge University Press.

D'Ambrosio, C., and E. Wolff. 2001. *Is wealth becoming more polarized in the United States? Working paper 330*. New York: Levy Economics Institute, Bard College.

Davidson, R., and J.-Y. Duclos. 2000. Statistical inference for stochastic dominance and for the measurement of poverty and inequality. *Econometrica* 68: 1435–1464.

Dinardo, J., and T. Lemieux. 1997. Diverging male wage inequality in the United States and Canada 1981–1988: Do institutions explain the difference? *Industrial and Labour Relations Review* 50: 629–651.

Dobkins, L., and Y. Ioannides. 2000. Dynamic evolution of the U.S. city size distribution. In *Economics of cities*, ed. J.M. Huriot and J.F. Thisse. Cambridge: Cambridge University Press.

Duclos, J.-Y., J. Esteban, and D. Ray. 2004. Polarization: concepts, measurement, estimation. *Econometrica* 72: 1737–1773.

Durlauf, S., and D.T. Quah. 1999. The new empirics of economic growth. In *Handbook of macroeconomics*, ed. J.B. Taylor and M. Woodford, vol. 1A. Amsterdam: North Holland.

Esteban, J.-M., and D. Ray. 1994. On the measurement of polarization. *Econometrica* 62: 819–851.

Foster, J., J. Greer, and E. Thorbecke. 1984. A class of decomposable poverty measures. *Econometrica* 52: 761–766.

Foster, J., and M. Wolfson. 1992. *Polarization and the decline of the middle class: Canada and the US Mimeo*. Nashville: Vanderbilt University.

Good, I., and R. Gaskins. 1980. Density estimation and bump-hunting by the penalised likelihood method exemplified by scattering and meteorite data. *Journal of the American Statistical Association* 75: 42–56.

Hartigan, J., and P. Hartigan. 1985. The dip test of unimodality. *Annals of Statistics* 13: 70–84.

Jenkins, S. 1996. Recent trends in the U.K. income distribution: What happened and why? *Oxford Review of Economic Policy* 12: 29–46.

Jones, C. 1997. On the evolution of the world income distribution. *Journal of Economic Perspectives* 11(3): 19–36.

Keefer, P., and S. Knack. 2002. Polarization, politics and property rights: Links between inequality and growth. *Public Choice* 111: 127–154.

Koshevoy, G., and K. Mosler. 1997. Multivariate Gini indices. *Journal of Multivariate Analysis* 60: 252–276.

Levy, F., and R. Murnane. 1992. US earnings inequality: A review of recent trends and proposed explanations. *Journal of Economic Literature* 30: 1333–1381.

Linton, O., Maasoumi, E. and Whang, Y.-J.2002. Consistent testing for stochastic dominance: A subsampling approach. Discussion Paper No.EM/02/433.STICERD, London School of Economics.

McFadden, D. 1989. In *Testing for stochastic dominance in studies in the economics of uncertainty (in honor of Josef Hadar)*, ed. T. Fomby and T. Seo. New York: Springer.

Quah, D. 1997. Empirics for growth and distribution: stratification, polarization, and convergence clubs. *Journal of Economic Growth* 2: 27–59.

Wang, Y.Q., and K.Y. Tsui. 2000. Polarization orderings and new classes of polarization indices. *Journal of Public Economic Theory* 2: 349–363.

Wolfson, M. 1994. When inequalities diverge. *American Economic Review Papers and Proceedings* 84: 353–358.

Wolfson, M. 1997. Divergent inequalities: Theory and empirical results. *Review of Income and Wealth* 43: 401–421.

Yitzhaki, S. 1994. Economic distance and overlapping of distributions. *Journal of Econometrics* 61: 147–159.

Poles of Development

N. Hansen

The term 'development pole' was first introduced by François Perroux (1955), who argued that analyses of economic development should concentrate on the processes by which various economic activities appear, grow in importance, and, in some cases, decline or disappear. Like Schumpeter, Perroux maintained that entrepreneurial innovation is primarily responsible for the development process, which involves a succession of dynamic sectors, or poles, over time. Although Perroux emphasized relations among industrial branches, the implications of the development pole notion have been elaborated mainly in terms of the geographic location of population and economic activities (Boudeville 1972; Hirschman 1958; Myrdal 1957).

The concept of geographic development poles, or growth centres, gained particular prominence in the context of the balanced versus unbalanced growth controversy of the late 1940s and 1950s. A number of economists held that economic development would best be accelerated by the simultaneous balanced growth of many interdependent undertakings. The principal rationale was that investments in both directly productive activities and infrastructure that would not be profitable in isolation would become profitable for the ensemble because of mutually beneficial external economies. However, the applicability of this strategy to newly developing countries was properly questioned on the ground that the resources required for carrying it out would be so great that a country disposing of such resources would not be underdeveloped in the first place. Critics of the balanced growth approach further pointed out that economic development does not in fact appear simultaneously and uniformly throughout an economy. Hirschman (1958) in particular maintained that development strategies for developing countries should concentrate on a few sectors rather than attempt to do too much at once with very scarce resources. In his view, development is communicated from leading sectors to the followers, from one firm to another. The advantage of this phenomenon over balanced growth, where every activity expands in step with every other, is that it leaves considerable scope to induced investment decisions, and therefore economizes the principal scarce resource, namely, genuine decision-making. However, Hirschman recognized that investments may well become overconcentrated in one or a few large cities because their external economies tend to be overrated by investment decision makers in the belief that nothing succeeds like success. Nevertheless, he believed that in the long run public investments would cease to be so concentrated in primate cities, because of national equity and unity considerations. In this regard he was clearly overly optimistic. Finally, Hirschman suggested that while infrastructure investments may be indispensable for the development of lagging regions, this would still represent only a permissive inducement mechanism. The essential task is the provision of continuously inducing activities in industry, agriculture and services.

During the 1960s it was widely held that, as a result of Keynesian macroeconomic policies, the economies of industrialized nations could continue to experience steady growth with relatively low unemployment and inflation rates. At the same time there was increased interest in structural problems that persisted despite the favourable aggregate context. For example, the growth of large metropolitan areas and the concomitant decline of some relatively peripheral regions became a concern in many industrialized countries, as well as a continuing concern in newly developing countires. It was frequently alleged, if not proven, that large urban agglomerations were too big, in the sense that the marginal social costs of further growth outweighed the marginal social benefits. Yet such places continued to grow because new entrants benefited from economies of agglomeration but did not bear the full costs associated with their entry. Critics argued that under these conditions, congested large cities and the nation as a whole would benefit if the growth of population and economic activity could be diverted to mediumsize development poles, whose accelerated growth could be induced by government policies with respect to infrastructure, taxation, capital subsidies and similar incentives. Proponents of this development pole strategy emphasized the advantages of economies of agglomeration in a relatively few urban centres and argued against policies that would spread development outlays too thinly over the national territory. The induced development poles would thus be economically efficient counter-magnets to the spontaneous development poles deemed to be too large. It was further argued that the induced development poles would generate beneficial ‘spread effects’ to their surrounding hinterland areas, so that in the long run the entire national territory would be characterized by ‘balanced growth’. An example of this strategy was the French spatial development policy of the 1960s, which designated eight metropolitan areas whose favoured growth would, it was hoped, counteract the growth of the Paris region.

The development pole strategies that were adopted in the 1960s fit in well with the hierarchical diffusion of innovations paradigm popular at the time. This approach to spatialtemporal development processes maintains that there are two principal features that characterize the spatial organization of economic activities: (1) a hierarchical system of cities, arranged according to the number and quality of functions performed by each city; and (2) a corresponding set of urban spheres of influence (urban fields) surrounding each of the cities in the system. Within this framework, development-inducing innovations are transmitted simultaneously in three ways: (1) outward from one or a few dominant national metropolitan ares to major regional urban centres; (2) downward from higher-order to lower-order cities in the urban hierarchy, in a pattern of hierarchical diffusion; and (3) outward from urban centres into their surrounding hinterland areas, that is, through radiating spread effects. The hierarchical diffusion of innovation paradigm is essentially a top–down model of development because it places considerable emphasis on continuing innovation adoption in the largest cities as the critical element for the subsequent extension of development over the entire urban-economic system.

Given this general setting, the role that induced development poles play in regional development can be regarded as a particular case of the general process of innovation diffusion. More specifically, development pole policies can be introduced if diffusion mechanisms are perceived to be operating too slowly, if ‘cumulative causation’ leads to increasing regional income disparities rather than to their reduction, or if institutional or historical barriers impede diffusion processes. The purpose of spatially selective public investments in development poles would be to hasten the focused extension of development to lower echelons of the urban hierarchy in peripheral regions, and to link the development poles more closely to the national urban system via higher-echelon cities in the urban hierarchy. It should be remarked that the innovation diffusion justification for a development pole policy does not deal with the issue of the actual or potential social costs of very large cities. Such places are viewed in an exclusively positive light, as the seedbeds of innovation, or at least the initial adopters of innovations conceptually generated elsewhere.

In retrospect, what have development pole policies accomplished? This is difficult to evaluate because it is hard to find any example of a development pole policy that has been vigorously implemented in practice. In many countries, development pole strategies have not really passed the stage of paper plans, and in many others the resources committed to implementing the plans have been too few to represent a genuine test of the strategy. Yet another pervasive problem has been the political difficulty of being selective in the choice of geographic development poles. Policies that have begun by attempting to concentrate investments in a relatively few urban centres have been diluted over time by pressures to include ever more centres, thus precluding the inducement of extensive economies of agglomeration in any one place.

In addition to the foregoing problems, the context for regional development policies has altered considerably over time, and especially since the mid-1970s. Here it is instructive to distinguish between the advanced industrial countries and the newly developing countries.

In the industrialized nation context, mounting evidence indicates that the hierarchical diffusion paradigm cannot be supported empirically. Development-inducing linkages run not only from larger to smaller cities, but also in the reverse order as well as between cities of similar size. Moreover, the notion that induced development poles will in turn induce development in their respective hinterlands has been undermined by evidence that interfirm and intersectoral linkages for the most part involve relatively distant locations. From the viewpoint of regional development policy, the problem of how to create local linkages has yet to be resolved. Although large city size is associated with technological progress in the hierarchical diffusion paradigm, there is no evidence that this is necessarily the case. Even in broadly regional terms, new industrial sources of innovation increasingly are widely dispersed. Finally, in many countries the aims of development pole strategies have tended to be realized since 1970, though few would attribute this phenomenon to such policies. Very large metropolitan areas have been declining in population or else have experienced much slower rates of growth than in the past; and many once stagnant or declining peripheral areas have experienced at least a modest degree of population increase and economic revival. In general, then, while spontaneous development poles continue to emerge outside of older industrial regions, the impetus to formulate deliberate policies to promote induced development poles has receded.

Development pole policies in newly developing countries have taken a number of forms. Some have concentrated on infrastructure in order to provide a critical minimum level of power, water, transportation and other public overhead facilities. Others have been based on the intermediate or heavy manufacturing activities of public enterprises; these projects have typically involved industrial complexes organized around such sectors as iron and steel, aluminium, petrochemicals and heavy engineering. What all these efforts have had in common is emphasis on the direct use of large-scale investment resources to generate structural changes through accelerated economic growth. However, the selection of development pole locations has typically been based on urban population growth projections and/or on national sectoral projections, but not on the development potential or demand of the surrounding hinterland areas. Consequently, polarized development under-takings have had only a very limited impact on their surrounding areas because the linkages involved in the development process have been largely with distant suppliers and markets, and because the derived demand for labour and for agricultural outputs has often stimulated migration and supplies from outside the regions where development poles haver been located. In addition, the highly capital-intensive nature of development pole activities has generated relatively low levels of employment in view of the considerable total resources invested.

In general terms, the principal criticism of polarized development strategies as applied in developing countries has been their failure to improve the social and economic well being of the large numbers of poor persons who live in rural peripheral regions. In recent years there has been a broadly based reaction against ‘top–down’ development efforts in favour of ‘bottom–up’ approaches that emphasize highly divisible, labour-intensive technologies applied to agriculture and to small and medium-size enterprises with direct linkages to agriculture and to rural and small town markets.

Despite criticisms of development pole policies, they are still being formulated in some countries, including Mexico and South Korea, where in each case decentralization away from the large, congested national capital is a major national objective. The evidence suggests that if such policies are to be successful they need to be broadened to include political, social and institutional changes as well as sectoral measures. Induced development poles need to be placed within a larger human settlement system framework, and human resource development policies need to be integrated with spatial-sectoral policies. The dissipation of scarce resources should be avoided by greater selectivity in location choices, and measures need to be taken to reduce the enclave nature of development poles. And political will needs to be sustained in the context of sufficiently long planning horizons.

In brief, then, prevalent attitudes toward development pole strategies have passed from an initial phase of optimism, to one of pessimism, to an emerging broader perspective that would include induced development poles as but one aspect of more comprehensive development planning (Hansen 1981).

See Also

- ▶ Location of Economic Activity
- ▶ Regional Development

Bibliography

Boudeville, J. 1972. *Aménagement du territoire et polarisation*. Paris: Génin.

Hansen, N. 1981. Development from above: The centre-down development paradigm. In *Development from above or below?* ed. W.B. Stöhr and D.R. Fraser Taylor. New York: John Wiley and Sons.

Hirschman, A.O. 1958. *The strategy of economic development*. New Haven: Yale University Press.

Myrdal, G. 1957. *Rich lands and poor*. New York: Harper and Brothers.

Perroux, F. 1955. Note sur la notion de pôle de croissance. *Economie Appliquée* 8, Series D, January– June.

Policy Reform, Political Economy of

Sharun W. Mukand

Abstract

Policymakers face political constraints that make enacting reform difficult. Since the late 1980s economists have developed a framework to analyse the deeper political underpinnings of policy inefficiency. This article develops a framework for delineating the key findings of this literature. It then briefly sketches out the role of institutions in facilitating policy reform.

Keywords

Adjustment costs; Commitment; Compensation; Democracy; Distributional conflict; Government failure; Institutional design; New political economy; Policy persistence; Policy reform; Public choice; Rational choice political economy; Rent seeking; Reputation; Stabilization policy; Time inconsistency

JEL Classifications

H0

Good policymaking is difficult. Inefficiencies abound in all areas of policymaking, due to constraints faced by the policymaker, be they informational, administrative or political. The subject of the political economy of policy reform is concerned with the *political* factors that make it difficult to reform policies and institutions. This field has focused on examining the impact of different political institutions on exacerbating or alleviating the ability to carry out reform, the consequences of these political constraints on policy outcomes as well as normative issues of institutional design that impinge on a policymakers' choice of policy.

The systematic exploration of the political economy underpinnings of policy reform began with two developments. First, there was an attempt to understand why governments in many developing countries failed to reform policies and institutions, despite low growth and stagnation and overall inefficiency (Rodrik 1996). Second, there were new developments in rational choice political economy. In particular, there was growing recognition of the power of the public choice critique of traditional policy analysis due to Buchanan and Tullock (1962). This literature emphasized that policymakers' preferences may be quite distinct from those of social planners and result in inefficient policy choices – that is, government failure. Therefore, policymakers' choices may be driven by a desire to retain office or may give very different weight to the preferences of some individuals (or groups) than might those of social planners. However, the insights from the public choice tradition were not explicitly grounded in a rational choice framework. This is where the literature on time-inconsistency spawned by the classic contribution of Kydland and Prescott (1977) played a crucial role. This literature demonstrated the importance of clearly specifying the policymaker's objectives and the constraints within a framework of optimization. However, the study of the political underpinnings of policy reform took off when insights from this 'new' political economy were used to deepen our understanding of economic crises, poverty traps and institutional inefficiencies in developing and transition economies.

Policy reform is difficult to achieve. Indeed, understanding the persistence of inefficient policy choices has been one of the central themes of much of the literature on policy reform. We can delineate the mechanisms proffered by the literature by focusing on two kinds of conflicts that make all policy reform more or less difficult. The first is the distributional conflict between different groups of citizens and individuals, be it due to differences in income, occupation, ethnicity or even religion. Given that much of policymaking is an attempt to balance these competing interests, the ability of a society to resolve this conflict is likely to affect its ability to reform. Second, the ability of a society to reform a failed policy may be due to conflict of interests between the politician–policymaker and those of the public. Institutions lie at the heart both these conflicts. Therefore, much recent work on the political economy of reform has focused on the interaction between political institutions, policy choices and inefficiency.

Distributional Conflicts and Policy Inefficiency

During the 1980s Latin America witnessed a number of macroeconomic crises due to delays in the enactment of any stabilization policy (Rodrik 1996). The puzzle was, why were these stabilizations delayed? In a near classic in this field, Alesina and Drazen (1991) addressed this issue. They showed that policy reform can be delayed due to a 'war of attrition' between two groups. Given uncertainty about the other group's willingness to bear a disproportionate burden of the adjustment costs, each group delays adjustment in the hope that the other caves in first. The economic crises worsens before one of the sides caves in and reform takes place.

At its broadest level inefficiency in policymaking in democracies arises from a commitment problem. Governments, which are vulnerable to losing power, are unable to commit to future policy outcomes. This failure in commitment can result in inefficient policy choices due to a variety of reasons (see Besley and Coate 1998). In particular, most policy reforms have distributional consequences, resulting in winners and losers. However, there is a time-inconsistency problem with promises of future compensation (see Acemoglu 2003; Robinson 1998, for an elaboration). Therefore, what is key is the inability of a government to credibly commit to compensate losers from economic reform. Not surprisingly, if losers are in a majority (or politically influential) they will be not only opposed but will be able to prevent the implementation of such a policy reform – even if it is efficient. Now, if a government through some form of taxes and transfers could credibly commit to compensate losers for their losses, then policy reform would be much easier to achieve. In part, the difficulty in making credible promises to compensate losers is that the gains and losses from policy choices are spread out over time, while the winners may not have enough resources to compensate the losers up front for their subsequent losses (see Dixit and Londregan 1996).

However, inability to compensate losers is not in itself sufficient for there to be a failure to adopt policy reform. If individuals are risk neutral, then they may well be willing to adopt a policy which has winners and losers. For example, consider an economy where 100 risk-neutral voters face the prospect of voting for or against a policy reform. If enacted this policy reform will result in 51 winners, each of whom gains five dollars, and 49 losers, each of whom stands to lose one dollar. We may suspect that since (in expected terms) all individuals stand to gain from the adoption of this policy reform, it will always be adopted, whether or not winners compensate the losers. In an important contribution, Fernandez and Rodrik (1991) suggest that this is not the case. They argued that, even in a world with risk-neutral agents, individual specific uncertainty about the identity of winners and losers from a reform may prove to be crucial. In particular, to continue with our example, consider the case where the identity of 49 of the 51 winners is common knowledge. In this case there is individual-specific uncertainty amongst the remaining majority about their identity as a winner or a loser. Observe that this uncertain majority has a negative expected payoff from the reform and will vote it down. Therefore under individual- specific uncertainty a majority may vote against a policy, despite the fact that a majority stands to benefit from it. In an extension, Jain and Mukand (2003) show that policy reform may fail to get enacted, despite the existence of tax-transfer compensation instruments.

Therefore, social conflict across groups coupled with a failure of credible and efficient means of conflict resolution result in a persistence of inefficient policies. In contrast to the above, the other class of models in this literature has focused on the agency conflict between the politician and the citizen.

Political Losers, Agency and Policy Inefficiency

Once in power, politicians earn both economic and non-economic rents. As such there may be a failure to enact any policy reform if it adversely affects the rents earned by the current incumbent politician. A number of mechanisms have been studied.

The prospect of earning rents from a status quo policy can make the adoption of policy reform by the politician much more difficult. Coate and Morris (1999) show that the mere introduction of a policy encourages the affected parties to make investments that increase their willingness to pay for retaining these policies in the future. If, in the future, the efficient policy is no longer the status quo policy then there may be a problem. Any government attempting to reform the status quo policy is likely to be vulnerable to lobbying by the now entrenched firms.

Indeed, in the presence of uncertainty about the policy choice, this inefficiency is exacerbated. For instance, many commentators wondered why US President Johnson persisted with military escalation in Vietnam though it was apparent to him (and most others) that such a policy was unlikely to work. Similarly, many analysts have been puzzled by the persistence of policymakers in many Latin American countries with extreme neoliberal policies, despite the fact that they seemed to not work. Majumdar and Mukand (2004) suggest that the reason is perhaps reputational. In particular, suppose that the initial choice of policy is a function of the policymaker's ability. In this case even if the policy seems to be failing, the policymaker may persist with it even if it is not efficient to do so. This is because a policy reversal by the incumbent will call into question the incumbent's competence in choosing this course of action in the first place. Fear of the adverse reputational (and electoral) consequence that such a policy reversal entails results in inefficient policy persistence.

We have delineated above a number of political mechanisms that make policy reform difficult. In response to the difficulty of reforming policies, recent work has focused on the implications of institutional innovations.

Institutions and Policy Reform

A country's institutional structure affects the political context of policymaking. Any inefficiencies in policy reform are likely to arise from an inability of existing institutions to mediate and resolve conflicts between groups or politicians. This central point about the role of institutional structures in stifling policy reform and growth has been made by Dixit (2004), Rodrik et al. (2004) and Acemoglu et al. (2004) among others. Accordingly, recent work has focused on issues of institutional design and the appropriate intervention most likely to change political (and economic) outcomes.

Broadly described, there have been two kinds of institutional interventions that can potentially alter the political equilibrium. The first is directed at resolving policy inefficiency arising out of distributional conflicts. The classic institution that facilitates conflict resolution is of course democratization and regular elections. While democratic elections bring their own inefficiencies, they have a positive first-order effect in that they give the electorate an opportunity to replace a policymaker who fails to undertake policy reform. Alternatively, a constitutional change in the nature of local government can help resolve policy inefficiencies. For instance, explicit political reservations for women and disadvantaged groups can directly alter an existing political equilibrium to one where policy reforms that benefit these groups will be more likely to take place (see Duflo and Chattopadhya 2004). The second kind of institutional intervention is one that helps limit the possibility of rent- seeking activity by the government. This may involve insulation of some of the policymaking apparatus from political pressures (as in an independent judiciary or central bank) or it may involve term limits or greater decentralization.

Final Remarks

Since the late 1980s the study of the political economy of policy reform has become an active area of research in political economy. Indeed, many of the seminal contributions to the area of the 'new political economy' were first made in an attempt to understand policy inefficiencies in developing countries and their failure to reform. Many empirical puzzles such as the inefficient delay in enacting reforms and the reversal of optimal policies have been addressed. In addition, the study of policy reform has given the initial impetus to the literature on institutions and underdevelopment.

However, much remains to be done. At an empirical level, we need to understand what institutions are likely to facilitate reform and prevent inefficiency. The need for micro-based studies is particularly important given the vast differences in history and socio-cultural norms that may result in both economic and political markets functioning in unexpected ways. Furthermore, more work needs to be done to understand the political economy of institutional change and its impact on policy reform. In particular, the role of leadership in catalysing institutional change and policy reform is poorly understood.

See Also

- ► Political Competition
- ► 'Political Economy'
- ► Political Institutions, Economic Approaches to
- ► Public Choice

Bibliography

Acemoglu, D. 2003. Why not a political Coase theorem? Social conflict, commitment and politics. *Journal of Comparative Economics* 31: 620–652.

Acemoglu, D., S. Johnson, and J. Robinson. 2004. Institutions as the fundamental cause of long run growth. In *Hand book of economic growth*, ed. P. Aghion and S. Durlauf. Amsterdam: North-Holland.

Alesina, A., and A. Drazen. 1991. Why are stabilizations delayed? *American Economic Review* 81: 1170–1188.

Besley, T. 2006. *Principled agents: The political economy of good government*. Oxford: Oxford University Press.

Besley, T., and S. Coate. 1998. Sources of inefficiency in a representative democracy: A dynamic analysis. *American Economic Review* 88: 139–156.

Buchanan, J., and G. Tullock. 1962. *The calculus of consent: Logical foundations of constitutional democracy.* Ann Arbor: University of Michigan Press.

Coate, S., and S. Morris. 1999. Policy persistence. *American Economic Review* 89: 1327–1336.

Dixit, A. 2004. *Lawlessness and economics: Alternative modes of governance*. Princeton: Princeton University Press.

Dixit, A., and J. Londregan. 1996. The determinants of success of special interests in redistributive politics. *Journal of Politics* 58: 1132–1155.

Duflo, E., and R. Chattopadhyay. 2004. Women as policy makers: Evidence from a randomized policy experiment in India. *Econometrica* 72: 1409–1443.

Fernandez, R., and D. Rodrik. 1991. Resistance to reform: Status-quo bias in the presence of individual-specific uncertainty. *American Economic Review* 81: 1146–1155.

Grossman, G., and E. Helpman. 1994. Protection for sale. *American Economic Review* 84: 833–850.

Jain, S., and S.W. Mukand. 2003. Redistributive promises and the adoption of economic reform. *American Economic Review* 93: 256–264.

Kydland, F., and E. Prescott. 1977. Rules rather than discretion: The time inconsistency of optimal plans. *Journal of Political Economy* 85: 473–490.

Majumdar, S., and S.W. Mukand. 2004. Policy gambles. *American Economic Review* 94: 1207–1222.

Robinson, J.A. 1998. Theories of bad policy. *Policy Reform* 1: 1–46.

Rodrik, D. 1996. Understanding economic policy reform. *Journal of Economic Literature* 34: 9–41.

Rodrik, D., A. Subramanian, and F. Trebbi. 2004. Institutions rule: The primacy of institutions over geography and integration in economic development. *Journal of Economic Growth* 9: 131–165.

Political Arithmetic

Phyllis Deane

Keywords

Chalmers, G.; Colquhoun, P.; Davenant, C.; Demography; Graunt, J.; King, G.; Massie, J.; National income; Petty, W.; Political arithmetic; Political economy; Young, A.

JEL Classifications

B1

The term ‘political arithmetic’ predates the term ‘political economy’. It was coined by Sir William Petty, a founder member of the Royal Society, who – being a scientist by education and a government economic adviser by career choice – deliberately set out to apply the new scientific methodology of the 17th century to the practical economic problems of the modern nation state. For the leading spirits of the scientific revolution which reached a climax in the second half of the 17th century, the common article of faith was a belief in the unity of theory and practice, combined with a conviction that the first step in the advancement of human understanding in any sphere of knowledge – whether in astronomy or in chemistry or in industrial or social technology – was to lay a foundation of direct, empirical observations. To quote Bacon’s *Novum Organum*:

> The roads to human power and to human knowledge lie close together, and are nearly the same; nevertheless, on account of the pernicious and inveterate habit of dwelling on abstractions, it is safer to begin and raise the sciences from those foundations which have relation to practice and let the active part be as the seal which prints and determines the contemplative counterpart.

That was the inspiration which underlay the foundation of the Royal Society and allowed men like Graunt and Petty, Newton and Boyle, Flamsteed and Hooke, to feel themselves part of the same intellectual community. That was also the inspiration for the first exercises in political arithmetic.

Sir William Petty was a medical practitioner (as was his contemporary Locke, and Quesnay a century later), and his interest in economic problems had been stimulated in Ireland, to which he went in the early 1650s as physician to the Cromwellian army of occupation. There he persuaded the civil authorities, faced with the problem of consolidating the conquest by making an orderly distribution of forfeited lands, to give him the task of organizing a comprehensive land survey. It was on the basis of this massive research project that he wrote *The Political Anatomy of Ireland* in the 1670s. But by then he had already published his *Treatise on Taxes and Contributions* (1662), which contained a miscellany of sharp observations, incisive economic analysis and forthright policy advice, mainly focused on English problems of public finance. He had also written (during the 1665–7 conflict with Holland) an essay in a similar analytical mould, concerned with the practical problems of financing the war, and it was in this connection that he produced his first estimates of national income and wealth for England and Wales (published posthumously as *Verbum Sapienti*).

Most of Petty’s pamphlets on economic questions were circulated privately and published posthumously, for the second half of the 17th century was an age in which giving politico-economic advice to governments was a perilous occupation. However, the distinctive message running through these writings was the importance of basing public economic policies on systematically compiled empirical evidence and reasoned quantitative estimates of the nation’s human and material resources. In the 1670s he spelt out and developed this theme in his most path-breaking work, *Political Arithmetick*, subtitled: ‘A discourse concerning the extent and value of lands, people, buildings, husbandry, manufacture, commerce, fishery, artisans, seamen, soldiers, public revenues, interest, taxes, superlucration, registries, banks, valuation of men, increasing of seamen, of militias, harbours, situation, Power at sea, etc. As the same relates to every country in general, but more particularly to the Territories of His Majesty of Great Britain and His Neighbours of Holland Zealand and France.’ Written in order to rebut those commentators who were lamenting the nation’s economic decline, Petty’s *Political Arithmetick* was an explicit attempt to apply a Baconian methodology to economic analysis. According to his preface:

> The Method I take to do this is not yet very usual; for instead of using only comparative and superlative Words and intellectual Arguments, I have taken the course (as a Specimen of the Political Arithmetick I have long aimed at) to express myself in terms of *Number, Weight* or *Measure*; to use only Arguments of Sense and to consider only such Causes as have visible Foundations in Nature, leaving those that depend on the Mutable Minds,

> Opinions, Appetites and Passions of particular men, to the Consideration of others.

Petty concluded his preface in the self-consciously undogmatic spirit of the 'new science' by inviting other seekers after truth to confront his results with rational criticism and new data:

> I hope all ingenious and candid Persons will rectifie the Errors, Defects and Imperfections which probably may be found in any of the Positions upon which these Ratiocinations were grounded. Nor would it misbecome Authority itself to clear the Truth of those matters which private Endeavours cannot reach to.

Petty is sometimes credited with having founded the first 'school' of economic thought. But it would be more accurate to say that he had launched the first scientific research programme in political economy. Indeed, in carrying over a Baconian scientific ideology to an analysis of public financial issues, he was simply epitomizing the spirit of his age, for political arithmetic was not narrowly economic in its scope. It was his friend John Graunt, for example, a London draper, who in 1662 published a pamphlet entitled *Natural and Political Observations Made upon the Bills of Mortality* and who, by applying a logical technique of coordination and deduction to the limited vital statistics that had been collected for London over the preceding century, took the first steps in the formulation of the modern science of demography. Using the death returns and other data regularly published in the London Bills of Mortality, plus some personally assembled data for a few country parish records, Graunt made the first reasoned estimates of total population, not only for the metropolis, but also for the country as a whole, and even set up the first life table. Significantly, Graunt's election to the Royal Society was made within a month of the publication of the first edition of his pamphlet, and was strongly supported by Charles II who (according to Sprat, the first secretary and historian of the Royal Society) 'gave his particular charge to His Society, that if they found any more such tradesmen, they should be sure to admit them all, without any more ado'.

It was indeed Graunt rather than Petty who inspired Gregory King to produce his demographic estimates in the 1690s, for Petty's way with figures was somewhat impressionistic. In particular, his results were less likely to be systematically crosschecked against the results suggested by alternative data sources, or by different sets of assumptions, than was the case for either Graunt or King. By the same token, Gregory King's estimates of national income were more meticulously justified, more detailed, more internally consistent, and hence more credible in their delineation of the dimensions of the economy and in international or intertemporal comparisons than were Petty's. Graunt and King, that is to say, were both more sophisticated economic statisticians than Petty. On the other hand, it was Petty's imaginative and ambitious use of his estimates as a basis for economic analyses and policy prescriptions that earned him his reputation as the leading political arithmetician. It is doubtful whether King's estimates, for example, would have had more than an ephemeral currency had they not been so brilliantly applied in the course of the polemical analyses of Charles Davenant – an MP and a public official of some weight, who had held inter alia the posts of commissioner of the excise 1683–9, inspector general of exports and imports 1705–15, and secretary to the commission set up to negotiate the Union with Scotland.

The half-century following the Restoration was the golden age of political arithmetic, but as a method of economic analysis it failed to develop appreciably during the next two centuries. Petty's (or occasionally King's) estimates of national income were often cited by mercantilist pamphleteers without any attempt at updating. True, there were from time to time new aggregative valuations of the nation's total income or product designed to put into perspective a polemical argument relating to a particular sector. For example, in 1760 Joseph Massie updated King's 'Scheme of the Income and Expence of the several Families of England Calculated for the Year 1688', in order to establish a framework for his own estimates of excise tax incidence in a polemic robustly entitled *A Computation of the Money that hath been exorbitantly Raised upon the People of Great Britain by the Sugar Planters in one Year from January*

1759 to January 1760; shewing how much Money a Family of each Rank Degree or Class hath lost by that rapacious Monopoly . . . Similarly, Arthur Young, who was mainly concerned to prescribe for and defend the economic interests of the agricultural sector, made some reasonably careful and well-informed estimates of the nation's agricultural output in the course of his reports on his *Tours* of the northern and eastern counties of England, and associated these estimates with more casual calculations of value added in manufacturing, commerce and various other industries. Young, who had a deservedly high reputation as an agricultural economist, but was undistinguished as a general economist, was probably the last of the political arithmeticians in the original sense of the term. Certainly he was the last economist to write under that banner, which by the 1930s had been annexed by the demographers. His treatise, entitled *Political Arithmetic, Containing Observations on the Present State of Great Britain and the Principles of her Policy in the Encouragement of Agriculture, to Which is Added a Memoir on the Corn Trade*, was a commentary on current agricultural issues which incidentally summarized and reconsidered some of his earlier national product estimates originally published in the Tours.

No doubt because the new discipline of political economy that took shape in the late 18th and early 19th centuries took off in more theoretical, less Baconian directions, political arithmetic lost its capacity to attract innovative exponents. Perhaps Adam Smith gave the *coup de grâce* to the whole approach when he announced in *The Wealth of Nations* that he had 'no great faith in political arithmetick' – though he was not above borrowing some of the political arithmeticians' estimates when they served the purpose of his argument. The third (1787) edition of the *Encyclopedia Britannica* (which did not include an entry for political economy) contained a lengthy piece on political arithmetic, defined as 'the art of reasoning by figures upon matters relating to government, such as the revenues, number of people, extent and value of land, taxes, trade, etc. in any nation'. The explanation went on: 'These calculations are generally made with a view to ascertain the comparative strength, prosperity etc. of any two or more nations.' Most of the entry was devoted to citations from Petty, although it referred also to Davenant, King, Graunt, Halley (the astronomer who had constructed a life table) and to various contributors to the demographic debates which flared up in the second half of the 18th century – such as Brakenridge and Price. After the 18th century, however, political arithmetic ceased to rate an entry in the *Encyclopedia Britannica* in its own right, and even the entries on political economy failed to notice it as a substantial episode in the history of economic ideas.

True, the idea of quantifying the total national income or wealth did not die, and when decennial population estimates were introduced, from 1801 onwards, the bases for such calculations became less speculative. The significance of Petty's role was that he had broken new ground in setting his analyses and associated policy prescriptions within a framework of national aggregates on whose structural relationships and absolute magnitudes the nation's productive strength and taxable capacity evidently hinged. A long, if sporadic, stream of national income estimates was accordingly produced by diligent researchers following in Petty's or King's footsteps – usually in relation to questions of war finance or taxable capacity or comparative economic strength. At the turn of the century, for example, Pitt's plans to raise an income tax to finance the French war stimulated a flurry of national income estimates. At about the same time, George Chalmers put Gregory King's *Natural and Political Observations* in an appendix to the fourth edition of his bestselling *Estimate of the Comparative Strength of Great Britain* (1802); and the appearance in print (for the first time ever) of King's famous table of incomes by families inspired Patrick Colquhoun, then researching the poverty problem, to update King's 1688 results for his *Treatise on Indigence* (1806). Less than a decade later, Colquhoun carried his statistical enterprise even further (with the aid of the first two population censuses) by publishing elaborate and detailed estimates of national income and wealth for the United Kingdom as a whole. This set the stage for the subsequent stream 19th-century estimates of national income which began essentially with Pebrer in 1833 and ended with Mulhall's *Dictionary of Statistics* (1890). Most of these, however, were exercises in descriptive statistics rather than in economic analysis.

In effect, then, Petty's aggregative approach to quantifying and analysing the nation's resources fell into disuse among leading economists as the abstract notion of a self-equilibrating economic system gradually took precedence over the essentially political concept of the royal domain as the central object of economic analysis. What 17th- or early 18th-century economic advisers typically addressed themselves to were the practical problems of the nation state, and these were seen as analogous to the problems of managing a household. Petty, for example, had dedicated his *Political Arithmetick* to the king, because it was the royal domain whose resources he was endeavouring to assess, and its management problems that the quantification was designed to inform. No doubt it was inevitable that when economists founded their theories on the assumption that 'things will have their course' in the politicoeconomic as in the natural universe, and while the role of the state within the wider economic system was assumed to be constrained by the sheer futility of legislating against the 'laws of nature', there was little incentive to extend national income analysis beyond the rudimentary levels it reached in the golden age of political arithmetic. Accordingly, the analytical approach to the study of national income was largely ignored by economists until the middle decades of the 20th century when J.M. Keynes's macroeconomic theorizing revolutionized their discipline.

See Also

► King, Gregory (1648–1712)
► Petty, William (1623–1687)

Bibliography

Colquhoun, P. 1806. *Treatise on indigence*. London: J. Hatchard.

Colquhoun, P. 1814. *Treatise on the wealth, power and resources, of the British Empire*. London: Joseph Mawman.

Davenant, C. 1771. *The political and economic works of that celebrated writer Charles D'Avenant*, 2 vols. Ed. Sir Charles Whitworth. London.

Deane, P. 1955. Implications of early national income estimates for the measurement of long-term economic growth in the United Kingdom. *Economic Development and Cultural Change* 4 (November): 3–38.

Deane, P. 1956–7. Contemporary estimates of national income in the nineteenth century. *Economic History Review*, Pt I, 8 (1956): 339–354; Pt II (1957): 451–461.

Graunt, J. 1662. Natural and political observations upon the bills of mortality. Reprinted in *The economic writings of Sir Wm. Petty*, ed. C.H. Hull, vol. 2. Cambridge: Cambridge University Press, 1899.

Hogben, L. 1938. *Political arithmetic: A symposium of population studies*. London: Allen & Unwin.

King, G. 1936. *Two tracts by Gregory King*. Ed. G.-E. Barnett. Baltimore: Johns Hopkins Press.

Massie, J. 1760. *A computation of the money that has been exorbitantly raised upon the people of Great Britain by the sugar planters*. London.

Mathias, P. 1957. The social structure in the eighteenth century: A calculation by Joseph Massie. *Economic History Review* 10: 30–45.

Mulhall, M. 1890. *Dictionary of statistics*. London: G. Routledge & Sons.

Pebrer, P. 1833. *Taxation, power, statistics and debt of the whole British Empire*. London: Baldwin and Cradock.

Petty, W. 1899. *The economic writings of Sir William Petty*, 2 vols. Ed. C.H. Hull. Cambridge: Cambridge University Press.

Young, A. 1770. *A six months tour through the North of England*. Vol. 4. London: W. Strahan.

Young, A. 1771. *The farmer's tour through the East of England*. Vol. 4. London: W. Strahan.

Young, A. 1774. *Political arithmetic. Pt I*. London: W. Nicoll.

Young, A. 1779. *Political arithmetic. Pt II*. London: T. Cadell.

Political Budget Cycles

Allan Drazen

Abstract

Theoretical and empirical research on political budget cycles is surveyed and discussed. Significant political budget cycles are seen to be primarily a phenomenon of the first elections after the transition to a democratic electoral system.

Keywords

Asymmetric information; Competence; Democracy; Electoral cycles; Electoral systems; Fiscal manipulation; Fiscal policy; Fiscal transparency; Imperfect information; Moral hazard; Political budget cycles; Political expenditure cycles; Political revenue cycles; Rational voting; Separating equilibria; Signalling; Targeted public spending

JEL Classifications

E6; H5

Political budget cycles are cycles in some component of the government budget induced by the electoral cycle. More specifically, the term most often refers to increases in government spending or the deficit or decreases in taxes (including changes relative to long-term trends) in an election year which are perceived as motivated by the incumbent's desire for re-election for himself or his party. Though political budget cycles may be seen as just one type of political cycle in macroeconomic variables, most research on cycles in economic variables induced by elections now focuses on budget cycles, and it is useful to study such cycles independent of political cycles in economic activity (the political business cycle). The shift in focus is due in part to the lack of strong empirical evidence for the existence of a political business cycle in many countries.

In contrast to the literature on the political business cycle – where development of formal models preceded the bulk of empirical testing – much empirical research on political budget cycles is based not on explicit models but on more conceptual arguments, with sophisticated formal models being developed later to show how the existence of cycles could be consistent with rational voters. In this article, we first review the basic conceptual arguments and then the formal models before considering the empirical research. There are two key empirical questions. The first is whether political budget cycles in fact exist in a large number of countries. Recent evidence, discussed below, suggests that they do not on the aggregate budget level, except for new democracies. The second key question, which underlies the first, is whether manipulation of the budget is an effective tool in gaining votes. Though it is widely believed that deficit spending in an election year in general gains votes for the incumbent, empirical research does not support this view.

Basic Conceptual Arguments

There are two main (and contradictory) views of pre-electoral fiscal manipulation. One is that politicians may be expected to engage in such manipulation and that empirically it is widespread. A simple argument supporting this view is that voters like low taxes and high government expenditures, and vote for incumbents who provide them. Opportunistic incumbents will therefore use expansionary fiscal policy before elections to increase the probability of re-election.

However, this simple argument is inconsistent with rational, forward-looking voters who are aware of government budget constraints both at a point in time and intertemporally. Since the non-smooth paths of taxes and government expenditures implied by election-year deficits are presumably costly, voters should dislike deficits in general and especially those seen as electorally motivated. They would therefore not reward incumbents who engage in election-year manipulation. Hence, the alternative view is that voters (especially in developed countries) are 'fiscal conservatives' who punish rather than reward fiscal manipulation. Evidence, discussed in greater detail below, suggests that this is the case in developed countries with established democracies.

A second argument is that if voters respond to good economic conditions by being more likely to vote for the incumbent, he will use expansionary fiscal policy to try to manipulate macroeconomic outcomes and provide higher growth. Hence, expansionary fiscal policy will help an incumbent's re-election prospects. However, even if good economic conditions help an incumbent's chances of re-election, it is not clear that fiscal manipulation will be effective – politicians may have very limited ability to manipulate the economy successfully, both because of a lack of technical ability to time the expansion accurately enough to happen just before the elections and because, as discussed above, rational, well-informed voters should not support such policies.

A more sophisticated argument on why rational voters may respond to pre-electoral fiscal expansions is that they have imperfect information about candidates' abilities or about the environment, and that a fiscal expansion signals incumbent ability or some other characteristic which voters value, so that it is effective in gaining votes. This was first formalized in the work of Rogoff, which is summarized below.

An alternative is that, if voters do punish election-year deficits or spending increases (as the data indicate for developed countries), electoral manipulation takes the form of changes in the composition of the budget rather than in its overall level (or the overall deficit). This may take the form of increases in spending that voters as a whole favour at the expense of those types of spending that voters may be believed to like less (or are less visible), or the form of expenditures targeted at some voters at the expense of other voting groups who are seen as electorally less valuable.

Signalling Models

The Basic Competence Model

Formal modelling of the signalling role of a pre-election fiscal expansion under asymmetric information was introduced by Rogoff and Sibert (1988) and Rogoff (1990). The models are based on unobserved 'competence', that is, the ability to deliver more public goods for the same level of taxes. Hence, more competent policymakers can generate higher welfare and so are preferred by voters. Competence is correlated over time, so that a candidate who is believed by voters before an election to be more competent than average (the presumed competence of his randomly drawn challenger, who is unable to signal) is expected to be more competent than average after the election as well. Voters therefore rationally prefer a candidate who delivers higher expenditures before an election, since this is a signal of higher competence.

The basic ideas can be represented by a simple version of the model in Rogoff (1990). There is an election at the end of the first period, with the leader who is elected remaining in office thereafter. Voters will choose the leader on the basis of any information they gather in the first period. The utility of the representative voter as of period t may be represented by

$$\Gamma_t = \sum_{s=t}^{T} \beta^{s-t}(g_s + v(k_s)) + \eta_t \tag{1}$$

where g_s is public consumption and k_s is public investment. The function $v(.)$ is assumed to be increasing, concave and satisfying the Inada conditions on its first derivatives as k goes to zero or infinity. The term η_t is a random shock in the election period $t = 1$ such that the outcome is not known *ex ante* to the incumbent setting policy. The voter maximizes the expected value of utility by choosing a candidate in an election at the end of the first period.

The production of public goods is represented as follows. If a leader has an 'administrative ability' or 'competence' ε, he can produce public goods at time t according to:

$$\varepsilon = g_t + k_{t+1} \tag{2}$$

where it is assumed that ε is not directly observable. Investment k must be chosen one period in advance, so that it is not currently observable. Hence, if a voter observes a high value of g_t, he does not know whether this reflects high ability of the policymaker (high ε) or high current public consumption 'bought' at the expense of a cut in some other component of public spending (here, public investment) at some point in the future. This is meant to represent the basic inference problem a voter faces when he observes high government spending before an election – does high observable government expenditure represent fiscal manipulation, in the sense of implying that taxes will be raised or other programmes cut in the future, or does it represent the ability of the leader to provide more goods or services without cutting future goods services?

Potential leaders are assumed to differ in their unobserved ability. Suppose there are two possible levels of ε: ε^H and $\varepsilon^L < \varepsilon^H$, where ability ε_j is expected to persist after the election. Let the prior probability that $\varepsilon = \varepsilon^H$ be $0 < \rho < 1$. The voter's inference problem is to use an observation of g to try to infer the probability that the leader is high-ability, that is, to form a posterior $\hat{\rho}(g)$.

The utility of the incumbent leader is given by:

$$E_t\Gamma_t + \left(\chi + q\sum_{s=t+1}^{T}\beta^{s-t}\chi\right)$$

where χ is the value of holding office and q is the probability of being re-elected at the end of the first period. A key point is that a policymaker's utility depends both on social welfare (the first term) and on his own private payoffs (the second term). If it depended only on social welfare, incumbents would choose the socially optimal fiscal policy and there would be no signalling. If it depended only on private payoffs, low-ability incumbents would mimic whatever high-ability incumbents do and there would only be a pooling equilibrium with no signalling.

At the beginning of period 1, the incumbent observes his ε^j, sets g_1 and k_2 (where k_1 is predetermined). Voters then observe g_1 and f_1 and then vote at the end of the period for either the incumbent or a randomly drawn challenger (who cannot signal his competence, which is average expected competence $\overline{\varepsilon}$ given the prior ρ.) In subsequent periods, the elected policymaker chooses g_t and k_{t+1} to maximize social welfare, given his competence ε. This first-best solution is given by maximizing (1) subject to (2), yielding $k^* = v'^{(-1)}(1/\beta)$ and $g^*(\varepsilon^j) = \varepsilon^j - k^*$. (This would also be the solution in period 1 if voters knew the incumbent's ε.) Since higher-ability incumbents provide more public goods, and thus higher utility, voters prefer a high-ability incumbent to the challenger of expected ability $\overline{\varepsilon}$, but prefer the challenger to a low-ability incumbent.

Under asymmetric information (that is, when the representative voter does not observe the incumbent's ε before voting, or cannot infer it because of imperfect information about the components of the budget), a voter's beliefs about an incumbent's ability are conditioned on his observation of g_1. These beliefs can be summarized as the posterior probability $\hat{\rho}(g_1)$ the voter assigns to the incumbent being of ability ε^H conditional on the value of g_1 observed. Given the voters' rational voting rule, an incumbent has an incentive to appear to be of high ability.

The equilibrium is a separating equilibrium in which the level of spending reveals the incumbent's competence type. A high-ability incumbent will spend just enough so that the low-ability incumbent will not find it optimal to mimic him. (Since a high-ability incumbent can invest $\varepsilon^H - \varepsilon^L$ in k_2 for the same level of g_1, and since politicians care about social welfare, concavity of $v(k)$ implies that the high-ability type can cut back on k at a lower marginal cost to himself than the low-ability type can, the signal of raising g_1 is less costly for him to send.) The low-ability incumbent will choose the first-best solution for his type, namely, $g_1 = g^*(\varepsilon^L) = \varepsilon^L - k^*$. Since this reveals his type he loses the election almost certainly.

If the values of ε^H and ε^L are far enough apart, then the high-ability incumbent can signal his type by choosing his first-best $g^*(\varepsilon^H)$, which the low-ability type won't mimic. However, if ε^H and ε^L are sufficiently close, then a high-ability incumbent can signal his type only by choosing $g_1 > g^*(\varepsilon^H)$. With a continuum of ability types, then each type separates from the type immediately 'below' him by choosing a $g_1 > g^*(\varepsilon^j)$, except for the lowest-ability type who plays his first best. Hence, there is the general result that there will be a fiscal expansion in an election year relative to non-election years, not because voters are naive but because they are sophisticated.

Timing of Signals

A question often raised about election-year expansions as a signal of competence (or some other desirable characteristic of a politician) is why the signal should be sent just before an election, rather than earlier in the politician's term. The argument in this sort of model is that information about such characteristics evolves over time, so that there is new information to be signalled in the time period before an election. At the same time the desirable characteristic must have some persistence, so that its preelectoral value provides information about its post-electoral value. (Formally, Rogoff modelled this by assuming there was an election at the end of every other period, with ability ε assumed to be the sum of the current period and previous period's i.i.d. shock, that is, an AR(1) structure. Therefore, information signaled by g_t in period t before an election was relevant for the post-electoral period $t + 1$, but not for the subsequent election at $t + 2$. This makes the incumbent's choice problem for choice of g_t fairly simple.)

Observability of Fiscal Policy

A key ingredient of this type of signalling model focusing on competence is voters' inability to observe the overall level of spending or of the deficit, for otherwise they could perfectly infer his competence. The reliance of this result on voters' lack of information is consistent with Brender and Drazen's (2005a) empirical finding of no statistically significant aggregate deficit or expenditure cycle in established democracies, where voters may be well-informed about fiscal outcomes. Gonzalez (2002) and Shi and Svensson (2002) extend the Rogoff model to study the effect of transparency on the magnitude of fiscal cycles. The basic result is that the higher the degree of transparency, the lower is the amount of distortion away from the first best in the political budget cycle. Shi and Svensson include a similar measure of transparency. Shi and Svensson further argue that, while the proportion of uninformed voters – who may be influenced by fiscal manipulation – is initially large, it is likely to decrease over time, thus decreasing the magnitude of budget cycles. They create a measure of the availability of information and show that as voters become more informed the magnitude of the cycle decreases. A key innovation of Shi and Svensson (2002) is that the policymaker chooses fiscal policy before he knows his competence level, so that all 'types' choose the same level of expansion. That is, the model focuses on moral hazard rather than signalling, as the other models do. An implication is a cycle in the aggregate deficit.

Unobserved Politician Preferences

The argument that, with high transparency, political cycles in aggregate expenditures or deficits are likely to be weak or non-existent (combined with empirical evidence on the absence of political cycles in budget aggregates in countries where transparency is seen as high) has led to alternative signalling models. If voters are fiscal conservatives, election-year fiscal manipulation may take the form of changes in the composition of the budget with overall spending and deficits held constant. These compositional changes may be either in categories of expenditures or in expenditures or transfers targeted to some voters at the expense of others.

Drazen and Eslava (2005, 2006) argue that, if it is the composition of spending or transfers, rather than their overall level, that is manipulated for electoral purposes, rational voters may be trying to infer something other than (or in addition to) competence from election-year fiscal policy. Voters who are targeted before an election want to know whether they will be similarly favoured after the election. They therefore suggest that a key unobserved characteristic of an incumbent politician is his preferences over groups of voters or types of expenditure. As in the Rogoff competence models, these preferences have some persistence over time, so that a voter who believes that the incumbent favours him before the election rationally expects some similarity in the composition of expenditures after the election as well. A voter thus faces an inference problem – whether receiving high targeted expenditures before the election signals a greater weight of his group in the incumbent's objective function than other voters or non-targeted expenditures, or whether it signals simply how 'swing' his demographic group is, meaning how many votes the incumbent can raise by targeting his group with expenditures. In both papers, Drazen and Eslava show the existence of an equilibrium in which voters rationally respond to election-year expenditures and politicians allocate expenditure on the basis of this behaviour. Politicians increase spending targeted to electorally attractive groups before elections, while they reduce other types of expenditure to satisfy the no-deficit constraint. As mentioned, a key result is that electoral manipulation arises even with fully rational voters. Drazen and Eslava (2006) further show that even when voters know how 'swing' their group is a political cycle may still arise.

There are several key differences between competence as the crucial unobserved characteristic and the approach of Drazen and Eslava, where a politician's preferences are unobserved and spending is targeted to some groups of voters or types of expenditure at the expense of others. First, in the latter approach, manipulation may occur even without affecting the aggregate deficit, consistent with empirical findings discussed below. Second, electoral fiscal manipulation arises even if voters can perfectly monitor the fiscal choices of an incumbent. Finally, political budget cycles in the Drazen and Eslava models arise even if all politicians are equally able to provide public goods.

Empirical Studies of Political Budget Cycles

Empirical studies of political budget began with the work of Tufte (1978) for the United States, followed by numerous other empirical studies for both developed and developing countries, as summarized in Drazen (2001). Political budget cycles were widely believed to be strongest for developing countries.

More recently, a number of papers have argued that, while these cycles are stronger in developing countries, they characterize democracies at all levels of economic development, and even non-democracies. Shi and Svensson (2002) find that, in a large panel of both democracies and non-democracies over the period 1975–95, the government deficit rises significantly in an election year in both developing and developed countries. (They show that the effect is far stronger in developing countries, consistent with earlier studies.) The economic effect is significant for the sample as a whole, the fiscal surplus falling on average in their full sample by one half to one per cent in an election year, depending on the estimation method they use. Persson and Tabellini (2003) restrict their sample to a group of 60 democracies from 1960 to 1998. They find a political revenue cycle (government revenues as a percentage of GDP decrease before elections), but no political cycle in expenditures, transfers, or the overall budget balance across countries or political systems. They argue that the electoral system (proportional versus majoritarian) and the governmental system (presidential versus parliamentary) is a key determinant of the nature of the cycle across countries.

However, Brender and Drazen (2005a) argue that the political deficit cycle in democracies is a phenomenon of recently democratized countries, that is, are found to be statistically significant only in the first few elections after a country has made a transition from being a non-democracy to a democracy (which holds true whether or not the formerly socialist economies are included). It is the strong political budget cycle in these countries that accounts for the political budget cycle in larger samples including these countries. Once these countries are removed from the larger sample, the political fiscal cycle disappears. This is true in both developed and developing countries. Hence, the stronger results previously found for developing countries reflect the fact that new democracies comprise a larger fraction of developing than developed country democracies. The 'new democracy' effect also helps explain previous findings of a stronger political cycle in weaker democracies (new democracies are a larger fraction of 'weak' than 'strong' democracies, with no significant cycle found in weak, old democracies). They also find that helps account for differences in the political cycle across government or electoral systems.

There is also a significant political expenditure cycle in the new democracies, with the very similar positive coefficients on the fiscal deficit and on expenditures in the analogous equations, while there does not appear to be a statistically significant revenue cycle. The deficit cycle in the new democracies thus appears to be driven by higher election-year expenditures.

Brender and Drazen suggest several explanations for their ‘new democracy’ finding. One is that fiscal manipulation may be used in new democracies because voters are inexperienced with electoral politics or may simply lack the information needed to evaluate fiscal manipulation that is produced in more established democracies. This suggests one way to reconcile the two contradictory views of preelectoral manipulation. The argument that politicians may be expected to engage in such manipulation may apply to new democracies, where it is possible to carry out such manipulation. The alternative that voters punish fiscal manipulation is applicable to established democracies, where voters have the ability to identify fiscal manipulation and punish such behaviour, so that politicians avoid it.

This is consistent with work by Gonzalez, Shi, and Svensson, discussed above, that focuses on information asymmetries in explaining budget cycles when voters are not naive. It is also consistent with findings by Akhmedov and Zhuravskaya (2004), who find similar evidence in regional elections in Russia after its transition to democracy. Using monthly data between 1996 and 2003, they found sizable but short-lived political budget cycles in local fiscal spending, which became significantly smaller over time and disappeared for most (but not all) fiscal instruments after two rounds of elections. Akhmedov and Zhuravskaya (2004) find similarly that measures of the freedom of the regional media and the transparency of the regional governments were important predictors of the magnitude of the cycle. Alt and Lassen (2006a) find that in OECD countries higher fiscal transparency also lowers the magnitude of the electoral cycle.

The absence of political cycles in budget aggregates in established democracies as a group does not, however, mean there are no electoral effects on fiscal policy. Established democracies appear to be characterized by cycles in the composition of spending rather than cycles in its overall level. Several papers find evidence of electoral composition changes in government spending at the sub-national level, including the United States (Peltzman 1992), Canada (Kneebone and McKenzie 2001), Colombia (Drazen and Eslava 2005), India (Khemani 2004), and Israel (Brender 2003). Drazen and Eslava (2005) present a signalling model of composition cycles with rational voters where the unobserved characteristic of politicians is their preferences for different types of expenditure, specifically those types of expenditure that voters as a whole prefer.

A second possible explanation for the new democracy effect follows from the Brender and Drazen (2005b) finding that fiscal balance has no significant effect on the probability of re-election, a surprising finding given the existence of a political budget cycle in new democracies. The authors suggest that these two findings may be reconciled by the possibility that fiscal expansions in election years in new democracies do not represent an attempt to gain voter support for the leader but reflect expenditures incurred in an attempt to consolidate democracy. Democracy is often not ‘consolidated’ in new democracies, that is, it is not accepted unconditionally by all citizens. An election year may be an especially dangerous time for the existence of the democracy itself, and thus may be a time when leaders have to spend money to retain popular support for the democratic regime to prevent its overthrow or subversion and the return to an autocratic system. One might then observe higher expenditures and deficits in an election year, but without fiscal expansion necessarily gaining votes for the incumbent over the challenger.

The Effect of Deficits on Re-election

In contrast to the fairly extensive direct tests of overall macroeconomic performance on election outcomes in the literature on political business cycles, there are few tests of fiscal performance on election outcomes, primarily at the sub-national level. These include Peltzman (1992), Brender (2003), and Drazen and Eslava (2005), who examine the direct effect of fiscal performance on re-election at the state and local levels in a single country (the United States, Israel, and Colombia respectively), and find that voters punish – rather than reward – loose fiscal policies in general, as well as in election years.

The only large cross-country study is by Brender and Drazen (2005b), who look at the effects of fiscal performance on re-election in a sample of 74 democracies (comprising 350 election campaigns) over the period 1960 to 2003. They estimate probit regressions giving the probability of an incumbent's re-election as a function of macroeconomic and fiscal variables. They find no evidence that expansionary fiscal policy helps a leader to get re-elected; in fact, it is likely to reduce the chances of re-election. In developed countries, especially established democracies, deficits lower the probability of re-election, with an effect that is both statistically and economically significant. In developing countries, the effect of deficits on re-election is close to zero and is not statistically significant. While voters in developing countries may be more tolerant of an expanding budget deficit in election years, even in these countries voters do not reward election-year deficits at the polls. Brender and Drazen find no statistically significant difference between the effect of deficits that are created by higher expenditures and of those that are created by lower revenue, although in the developed countries the effect of revenue reductions (as a share of GDP) is somewhat larger.

They also find that in established democracies in developed countries voters punish election-year deficits and deficits over the incumbent's term of office. The effects are quite substantial quantitatively. An increase of one percentage point in the ratio of the central government surplus to GDP over the term can increase the probability of re-election by 3–4.5 percentage points in the developed, established democracies, and an increase of one percentage point in the surplus during an election year increases the probability of reelection by between seven and nine percentage points.

The Brender–Drazen results indicate that controlling for the type of political system (parliamentary versus presidential) or the type of electoral system (majoritarian versus proportional) does not change the effect of the election year deficit and growth, nor does whether elections were held at their scheduled date or early. Similarly, they find no significant effect of the level of democracy on the finding that deficits do not help re-election chances of an incumbent.

See Also

▶ Political Business Cycles

Bibliography

Akhmedov, A., and E. Zhuravskaya. 2004. Opportunistic political cycles: Test in a young democracy setting. *Quarterly Journal of Economics* 119: 1301–1338.

Alt, J., and D. Lassen. 2006a. Transparency, political polarization, and political budget cycles in OECD countries. *American Journal of Political Science* 50: 530–550.

Alt, J., and D. Lassen. 2006b. Fiscal transparency, political parties, and debt in OECD countries. *European Economic Review* 50: 1403–1439.

Brender, A. 2003. The effect of fiscal performance on local government election results in Israel: 1989–1998. *Journal of Public Economics* 87: 2187–2205.

Brender, A., and A. Drazen. 2005a. Political budget cycles in new versus established democracies. *Journal of Monetary Economics* 52: 1271–1295.

Brender, A., and A. Drazen. 2005b. How do budget deficits and economic growth affect reelection prospects? Evidence from a large cross-section of countries. Working Paper No. 11862. Cambridge, MA: NBER.

Drazen, A. 2001. The political business cycle after 25 years. In *NBER macroeconomics annual 2000*, ed. B. Bernanke and K. Rogoff. Cambridge, MA: MIT Press.

Drazen, A., and M. Eslava. 2005. Electoral manipulation via voter-friendly spending: Theory and evidence. NBER working paper 11085. Cambridge, MA: NBER.

Drazen, A., and M. Eslava. 2006. Pork barrel cycles. NBER working paper 12190. Cambridge, MA: NBER.

Gonzàlez, M. 2002. Do changes in democracy affect the political budget cycle? Evidence from Mexico. *Review of Development Economics* 6: 204–224.

Khemani, S. 2004. Political cycles in a developing economy: Effect of elections in the Indian states. *Journal of Development Economics* 73: 125–154.

Kneebone, R., and K. McKenzie. 2001. Electoral and partisan cycles in fiscal policy: An examination of Canadian provinces. *International Tax and Public Finance* 8: 753–774.

Peltzman, S. 1992. Voters as fiscal conservatives. *Quarterly Journal of Economics* 107: 327–361.

Persson, T., and G. Tabellini. 2003. *The economic effects of constitutions*. Cambridge, MA: MIT Press.

Rogoff, K. 1990. Equilibrium political budget cycles. *American Economic Review* 80: 21–36.

Rogoff, K., and A. Sibert. 1988. Elections and macroeconomic policy cycles. *Review of Economic Studies* 55: 1–16.

Shi, M., and J. Svensson. 2002. Conditional political budget cycles. Discussion Paper No. 3352. London: CEPR.

Tufte, E. 1978. *Political control of the economy*. Princeton, NJ: Princeton University Press.

Political Business Cycles

Allan Drazen

Abstract

Theoretical and empirical research on political business cycles, both opportunistic and partisan, is surveyed and discussed. The evidence for the existence of empirically significant opportunistic political business cycles is argued to be mixed.

Keywords

Central bank independence; Competence; Imperfect information; Infinite horizons; Inflationary expectations; Natural rate of unemployment; Opportunistic vs. partisan business cycles; Phillips curve; Political budget cycles; Political business cycles; Rational expectations; Rational voting; Signaling

JEL Classifications

E6; H5

Political business cycles are cycles in macroeconomic variables – output, unemployment, inflation – induced by the electoral cycle. (Political cycles in fiscal policy variables, termed 'political budget cycles', are treated in a separate article.) Key questions this literature addresses include the following. Are such cycles observed in the data? What are the political and economic mechanisms that lead to such cycles? What do they imply about voter behaviour?

There are two basic types of models. 'Opportunistic' political business cycles are expansions in economic activity induced by an opportunistic incumbent before an election meant to increase his chances of re-election. 'Partisan' political business cycles are fluctuations in macroeconomic variables over or between electoral cycles resulting from leaders having different policy objectives.

Opportunistic Models

Formal models of the opportunistic business cycle began to appear in the mid-1970s, the most influential of which was that of Nordhaus (1975). The structure of the economy is summarized by a downward-sloping Phillips curve, yielding a trade-off between unemployment and unexpected inflation. Inflation expectations are formed adaptively on the basis of past observed inflation. Identical voters base their voting decisions on aggregate inflation and unemployment outcomes relative to their most preferred outcomes. They have a preference for both low unemployment and low inflation, but, in evaluating incumbents on the basis of macroeconomic performance, they have short memories and no foresight. An opportunistic incumbent policymaker has no preferences over inflation and unemployment *per se* and cares only about re-election. The slow adjustment of inflation expectations to economic stimulation, combined with myopic voters, allows an opportunistic incumbent to manipulate macroeconomic time paths to his electoral benefit. He stimulates the economy before the election to reduce unemployment, with the inflationary cost of such a policy coming only after.

More formally, the basic opportunistic model may be simply represented as follows. The objective of the policymaker is to maximize his probability of re-election, where voting behaviour is retrospective in that it depends on economic performance under the incumbent in the past. Economic performance in a period is measured by the behaviour of current inflation π_t and unemployment U_t, so that voter dissatisfaction in any period can be represented by a loss function which is increasing in these two variables. Consider, for simplicity:

$$L(U_t, \pi_t) = U_t + \theta \frac{(\pi_t)^2}{2} \tag{1}$$

where θ is the relative weight the electorate puts on inflation deviations relative to unemployment and where (for simplicity of exposition) it is assumed that the representative voter's most preferred rate of inflation is zero.

One may then posit a retrospective voting function for an election at the end of period *t,* of the form:

$$V_t = \Psi\left(\sum_{s=0}^{T-1} \gamma(s) L(U_{t-s}, \pi_{t-s})\right) \qquad (2)$$

yielding the number of votes V_t for the incumbent as a decreasing function of loss from economic outcomes ($\Psi' < 0$). The exogenous length of time between elections is T periods, and $\gamma(s)$ is the weight voters put on a loss s periods in the past. $\gamma(s)$ is assumed to be decreasing in s, that is, past economic outcomes have a smaller effect on votes at t the further in the past they are. If $\gamma(s)$ is rapidly decreasing in s, very recent events are weighted most heavily. In the extreme, $\gamma(s) = 0$ if for $s > 0$, then only economic outcomes in the year of the election affect voting. The electoral mechanism is not made more specific. One could add a stochastic element to allow for the possibility of an incumbent losing the election.

In the Nordhaus model, the structure of the economy is summarized by an expectations-augmented Phillips curve relating the difference between the actual and the natural rates of unemployment U_t^N to the difference between actual and expected inflation π_t^e :

$$U_t = U_t^N - \left(\pi_t - \pi_t^e\right) \qquad (3)$$

To close the model one must specify the formation of expectations. Crucial to the main results of the above models is some form of backward-looking expectations, so that inflationary policy in an election period is not fully anticipated and can therefore lower the unemployment rate. A standard formulation of adaptive determination of the expected rate of inflation:

$$\pi_t^e = \pi_{t-1} + \alpha\left(\pi_{t-1}^e - \pi_{t-1}\right) \qquad (4)$$

where α is a coefficient between 0 and 1 representing the speed with which expected inflation adapts to past expectational errors. This may be solved to yield π_t^e as a weighted declining sum of past inflation rates.

This four-equation system may then be solved for unemployment and inflation over the electoral cycle. When voters have 'short memories' ($\gamma(s)$ small for $s > 0$) a political business cycle will emerge if the incumbent wants to maximize his probability of re-election. In the period immediately after the election the government engineers a recession via contractionary monetary policy to bring down inflationary expectations. The incumbent keeps economic activity low to keep expected inflation low until the period immediately before the next election, so that a given rate of economic expansion (induced by a monetary surprise) can be obtained at a relatively low rate of inflation. The government then stimulates the economy via expansionary monetary policy, unemployment falling due to high unanticipated money growth. The levels of monetary expansion and unemployment are those which maximize voter satisfaction in the election period. In the next election cycle the same behaviour is repeated, with contractionary monetary policy to bring down inflation expectations. Hence, the possibility of influencing the probability of re-election, combined with the structure of the economy, yields a cycle in economic activity which would not be present with a planner with an infinite horizon. The political cycle thus induces a cycle in economic activity and inflation.

Though these models capture the incentive for opportunistic policymakers to manipulate policy and the macroeconomic cycle that may result, a number of conceptual and empirical objections may be raised. First, incumbents running for re-election do not control monetary policy in countries with independent central banks. However, there is evidence that nominally independent central banks often accommodate the executive branch's pressures for monetary policy during election years in order to prevent sharp movements in interest rates (see, for example, Woolley 1984, for evidence for the United States). Hence, politically motivated monetary policy in an election year may be a good approximation to reality.

Second, one may question whether voters are really as unsophisticated as the basic models assume, both in the way they form expectations of inflation and in the way they assess government performance. Voters realize that 'election-year economics' may be used to win their votes and hence may be sceptical of an economic upturn in the months before an election. More formally, their expectations of inflation should take the possibility of an election-year monetary expansion into account (which would then nullify its effects since it is no longer a surprise). An intermediate view is that voters have less-than-perfect information about the causes of economic fluctuations and take good economic performance as indicating incumbent competence. Hence voting for the incumbent when times are good is consistent with rationality when voters have imperfect information. This has been argued by Nordhaus (1989) and has been formalized using signalling models, as discussed below.

Partisan Models

In partisan models, cycles are induced by differences among parties in their ideology and their economic goals. The basic partisan model is due to Hibbs (1977), based on different preferences over inflation and unemployment across parties. One replaces the voters' loss function (1) with one representing the preferences of a party j, for example,

$$L^j(U_t, \pi_t) = \frac{\left(U_t - \tilde{U}^j\right)^2}{2} + \theta^j \frac{\left(\pi_t - \tilde{\pi}^j\right)^2}{2} \qquad (5)$$

where $\tilde{\pi}^j$ is party j's target rate of inflation, $\tilde{U}^j$ is party j's target unemployment rate, and θ^j is the weight party j puts on deviations of inflation from target inflation relative to deviations of unemployment from target. The two parties, say a right-wing party R and a left-wing party L, are characterized, for example, by $\tilde{U}^L \leq \tilde{U}^R$, $\theta^L \leq \theta^R$, and $\tilde{\pi}^L \geq \tilde{\pi}^L$. Thus, the left-wing party will pursue a more expansionary monetary policy throughout its term. Using the same specification of the relation between unemployment and inflation as in (3) and a similar specification of backward-looking expectations (4), one may derive a cycle in which the level of economic activity and inflation varies with the ideology of the incumbent.

Rational Voters

Early models in both strands of the literature were often criticized in their modelling of expectations, since the backward-looking nature of expectations was crucial for some of the results. Hence, in both strands the focus has shifted to models in which voters form their expectations rationally, with the question being whether a political budget cycle will still exist with rational, forward-looking voters.

In the context of an opportunistic political budget cycle, the key argument is that some characteristic of policymakers is unobserved, and the voters' inference problem over an incumbent's 'type' will imply it is optimal to vote more heavily for the incumbent when economic outcomes are favourable. A leading unobserved characteristic is the incumbent's 'competence'. More competent policymakers produce better outcomes, and competence has some persistence over time. Therefore, good outcomes in the time period before the election may signal high competence of the incumbent (relative to a challenger who cannot signal), which is expected to persist after the election. Hence, when competence cannot be observed directly, it may be optimal for voters to vote more heavily for the incumbent if times are good.

This argument may be formalized in an imperfect information framework. The first formal models concerned political budget cycles in work by Rogoff (for example, Rogoff 1990). Persson and Tabellini (1990) and Lohmann (1998) present similar models of unobserved policymaker ability as applied to cycles in economic activity. High economic activity before an election signals a high-ability incumbent, that is, higher than the average expected ability of the challenger. Since ability has a persistent component, voters expect better economic performance from the incumbent than from the challenger after the election as well, and hence vote for him.

Alesina (1987) introduces rational expectations into the original partisan model of Hibbs, so that fluctuations in inflation and unemployment are driven by partisan differences combined with uncertainty about election outcomes. Close elections imply the sort of fluctuations Hibbs found, but because expansionary monetary policy by a left-wing policymaker (for example) is not fully anticipated before an election and therefore will lead to a fall in unemployment after the election. A key difference from the Hibbs model is that any effect on unemployment will no longer be present after inflation expectations are adjusted. Hence, the effects on unemployment will be concentrated early in a leader's term of office and disappear in the latter part of the term once the leader's preferences are known.

Empirical Testing

The existence of opportunistic political business cycles has been subject to extensive empirical testing. There are two key questions: are election years characterized by economic expansions? Do voters respond to 'good times'?

The standard test for the existence of a political cycle is to run an autoregression of an economic performance measure on itself, a small set of economic variables, and political dummies, that is, a regression:

$$Y_t = \sum_{i=1}^{s} a_i Y_{t-i} + b_0 + \sum_{j=1}^{k} b_j X_{jt} + d\,PDUM_t + \varepsilon_t \qquad (6)$$

where Y is an outcome variable such as output growth, the X_j are control variables, and PDUM is a political dummy variable (or set of variables) meant to represent a given political model. The autoregressive specification for Y_t is adopted as a parsimonious representation of the time series behaviour of Y, instead of using a structural model. The hypothesis that output growth, for example, is higher in election years would be represented by setting $PDUM_t$ equal to 1 in election years and zero otherwise, and testing whether the coefficient d is statistically significant.

The evidence for a political cycle in outcomes is quite mixed, with most studies finding little evidence of opportunistic political cycles in developed countries. Much of this evidence is summarized in Alesina et al. (1997) and Drazen (2000).

The evidence on voter response to economic conditions is also mixed. Generally, the effect of growth on re-election probabilities was found to be insignificant in most cross-section studies in developed countries (see Brender and Drazen 2005, for a summary). The United States seems to be an exception to these findings. The most influential paper on voter response in the United States is probably that of Fair (1978), who found that an increase in real economic activity in the year of the election, as measured either by the change in real per capita GNP or the change in unemployment in the election year, has a strong positive effect on the incumbent's vote total in US presidential elections. Alesina and Rosenthal (1995) find similar results.

Brender and Drazen (2005) confirm the insignificant effect of growth on re-election probabilities in developed countries in a large cross-section study of a sample of 74 democracies over the period 1960–2003. In contrast, they find that in less developed countries higher growth in real GDP has a positive and statistically significant effect on the probability of re-election. They then remove from the overall growth rate the part that voters might attribute to global developments and find that in the less developed countries it is the component of growth associated with domestic influences that accounts for the highly significant effect of growth on re-election, while the part attributable to global economic growth has no statistically significant effect on the probability of re-election. In the developed countries they find that neither the effect of global growth nor the effect of domestically induced growth is statistically significant.

There has been less empirical testing of the partisan political business cycle. The striking empirical regularity in the United States since the Second World War is that economic activity is substantially higher under Democrats than under Republicans in the first part of their four-year terms, but more similar in the second part of their terms, consistent with the Alesina model. However, Faust and Irons (1999) argue that the data do not give strong support to any partisan model. For the OECD, Alesina et al. (1997) find supporting evidence for the rational partisan model in a number of countries.

Overall, the focus of both theoretical and empirical research has shifted to political budget cycles, in large part due to the weak empirical evidence for the existence of an opportunistic political business cycle in many countries, combined with the widespread view that, nonetheless, election year manipulation of some sort is a common phenomenon.

See Also

▶ Political Budget Cycles

Bibliography

Alesina, A. 1987. Macroeconomic policy in a two-party system as a repeated game. *Quarterly Journal of Economics* 102: 651–678.

Alesina, A., and H. Rosenthal. 1995. *Partisan politics, divided government, and the economy*. Cambridge: Cambridge University Press.

Alesina, A., N. Roubini, and G. Cohen. 1997. *Political cycles and the macroeconomy*. Cambridge, MA: MIT Press.

Brender, A. and Drazen, A. 2005. *How do budget deficits and economic growth affect reelection prospects? Evidence from a large cross-section of countries*. Working paper no.11862. Cambridge, MA: NBER.

Drazen, A. 2000. *Political economy in macroeconomics*. Princeton: Princeton University Press.

Fair, R. 1978. The effect of economic events on votes for president. *Review of Economics and Statistics* 60: 159–172.

Faust, J., and J. Irons. 1999. Money, politics, and the postwar business cycle. *Journal of Monetary Economics* 43: 61–89.

Hibbs, D. 1977. Political parties and macroeconomic policy. *American Political Science Review* 71: 1467–1487.

Lohmann, S. 1998. Rationalizing the political business cycle: A workhorse model. *Economics and Politics* 10: 1–17.

Nordhaus, W. 1975. The political business cycle. *Review of Economic Studies* 42: 169–190.

Nordhaus, W. 1989. Alternative approaches to the political business cycle. *Brookings Papers in Economic Activity* 1989 (2): 1–49.

Persson, T., and G. Tabellini. 1990. *Macroeconomic policy, credibility, and politics*. London: Harwood.

Rogoff, K. 1990. Equilibrium political budget cycles. *American Economic Review* 80: 21–36.

Tufte, E. 1978. *Political control of the economy*. Princeton: Princeton University Press.

Woolley, J. 1984. *Monetary politics*. Cambridge: Cambridge University Press.

Political Competition

David Austen-Smith

Abstract

This article is limited to interaction between candidates and voters and examines the cases of two-candidate competition and multiple candidate competition. It employs the spatial model of elections introduced to study single-issue politics and generalized to study multiple-issue politics in order to explain the alternatives strategically offered to voters by candidates or parties competing for electoral office.

Keywords

Black, D.; Downs, A.; Duverger's Law; Electoral competition; First Welfare Theorem; Hotelling, H.; Median voter theorem; Mixed strategy equilibrium; Myopia; Nash equilibrium; Political competition; Rational choice models; Voting

JEL Classifications

D72

In its most general form, political competition concerns the struggle of ideas for organizing societies. This article, however, focuses explicitly on one concrete manifestation of this struggle, namely, electoral competition. Any convincing and general explanation of electoral competition must account for the role of money in campaigns, for the behaviour of interest groups and the implications of party organization. This article addresses only the interaction between candidates and voters. Although not the only framework for studying the topic, the spatial model of elections introduced by Hotelling (1921) and Downs (1957) for single-issue politics and generalized by Davis and Hinich (1966, 1967) to multiple-issue politics, is surely the most widely used. The principal goal of the theory is to explain the alternatives strategically offered to voters by candidates or parties competing for electoral office.

Two-Candidate Competition

A Benchmark Model

There are two candidates A,B, and a large finite set of voters $N = \{1, \ldots, n\}$. The policy space is a convex and compact set $X \subset R^k$, where $x \in X$ is a typical feasible policy. Each voter $i \in N$ has policy preferences on X representable by a continuous and strictly quasi-concave utility function $u_i : X \to R$; let $\mathbf{u} = (u_1, \ldots, u_n)$ denote the preference profile over X. Candidates too have preferences although they need not be defined directly on the policy space, X; candidates, for example, might plausibly be more interested in winning office than in policy per se, or in some combination of winning and the policy eventually implemented, irrespective of who wins. On the assumption of complete information on the part of voters regarding candidates' motivations and policy platforms, and on the part of candidates regarding voters' preferences, however, introducing policy motivations for candidates leads to no essential change in the predictions of the model with purely office-motivated candidates (Calvert 1985; Duggan and Fey 2005). Some implications of assuming that candidates are policy motivated when there is some uncertainty about payoffs are considered later; for now, suppose that each candidate is motivated solely by the desire to win office.

The election is for a single office and is determined by a plurality rule. On the assumption of no abstention, a voting strategy for any citizen $i \in N$ is a mapping $v_i : X^2 \to [0, 1]$, where $v_i(a, b)$ is the probability that i votes for candidate A when A chooses electoral platform $a \in X$ and B chooses a platform $b \in X$. Given a profile of vote strategies $\mathbf{v} = (v_i)_{i \in N}$, $V_j(a, b| \mathbf{v}) \in [0, n]$ is the expected number of votes cast for candidate j. Let $\Pi(a, b| \mathbf{v}) = V_A(a, b| \mathbf{v}) - V_B(a, b| \mathbf{v})$ denote A's expected plurality, so $-\Pi(a, b| \mathbf{v})$ is B's expected plurality. Then candidate j's payoff under plurality rule is 1 if her realized plurality is strictly positive, -1 if it is strictly negative, and zero otherwise.

The strategy space for each candidate is X. Maximizing j's payoffs in this setting is equivalent to maximizing j's expected plurality. Thus, A chooses $a \in X$ to maximize $\Pi(a, b| \mathbf{v})$ and B chooses $b \in X$ to minimize, $\Pi(a, b| \mathbf{v})$. Under the assumptions of no abstention and two candidates, maximizing $\Pi(a, b| \mathbf{v})$ is equivalent to maximizing $V_A(a, b| \mathbf{v})$. Later, I consider some implications of admitting abstention and multiple candidates where the equivalence fails. An equilibrium to the game is a vector of undominated strategies $(a^*, b^*, \mathbf{v}^*)$ such that each voter $i \in N$ is maximizing u_i conditional on (a^*, b^*) and the voting strategies of all $j \in N/\{i\}$ and, given $\mathbf{v}^*$,

$$\Pi(a^*, y| \mathbf{v}^*) \geq \Pi(a^*, b^*| \mathbf{v}^*) \geq \Pi(y, b^*| \mathbf{v}^*)$$

for all $y \in X$.

Existence of equilibrium in the model when candidates use pure strategies is a problem. The majority core, that is, the set of alternatives $x \in X$ such that no alternative is strictly preferred by a majority to x, is guaranteed to be non-empty only when the dimensionality of the policy space, k, is 1 (Plott 1967; McKelvey 1979; Schofield 1983) and it is not hard to see that the set of equilibria in pure candidate strategies coincides with the majority core. The most familiar example of this coincidence is the median voter theorem for $X \subseteq R$ (Downs 1957; Black 1958), predicting candidate convergence on the median most preferred policy. On the other hand, if a policy space of any finite dimension is approximated with a finite grid, irrespective of how fine the grid might be, the classical Nash equilibrium existence theorem implies the existence of an equilibrium in mixed candidate strategies. Furthermore, in the finite case with no majority indifference, the mixed strategy equilibrium is unique and symmetric (Laffond et al. 1993): both candidates adopt the same mixed strategy and thus, *ex ante*, the candidates are equivalent from a policy perspective, just as they are under the median voter theorem.

The difficulty in proving a general mixed strategy equilibrium existence result for the spatial voting model with a continuum of alternatives lies in the absence of sufficient continuity in the mapping that connects pairs of policy positions to vote shares: a small unilateral change in one candidate's position can result in the candidate's vote share changing from less to more than one-half of the electorate, thus inducing a discrete jump in her payoff. But these discontinuities often arise as a result of the presumption that indifferent individuals in the spatial model necessarily vote for each candidate with probability one-half. If this assumption is relaxed and individuals are restricted only to symmetric voting strategies, thus allowing the probability of indifferent individuals voting for one or other candidate to be sensitive to the platforms offered, then existence of a mixed strategy equilibrium is guaranteed (Duggan and Jackson 2004). Characterizing mixed candidate strategies, however, is not easy. McKelvey (1986) and Banks et al. (2002) provide some insight by showing that the support of any mixed strategy equilibrium (essentially) lies within the closure of the uncovered set: say that an alternative x is covered by an alternative y if x is strictly majority preferred to y and, further, that any alternative z that defeats x also defeats y; the uncovered set is then the set of alternatives that are not covered (Miller 1980). The uncovered set generally exists in the spatial model and, moreover, if a sequence of (continuous and strictly quasi-concave) preference profiles converges uniformly to a profile $\mathbf{u}^*$ at which the majority core is non-empty, then (loosely speaking) the associated sequence of uncovered sets converges to the core at $\mathbf{u}^*$. Thus, for any profile $\mathbf{u}$ 'close' to a profile $\mathbf{u}^*$ supporting pure strategy equilibria to the election, the realized policy platforms offered to the electorate at $\mathbf{u}$ are 'close' to the (pure strategy) equilibrium policies offered at $\mathbf{u}^*$.

Results on the uncovered set notwithstanding, a convincing interpretation of mixed candidate strategies in electoral competition is elusive. A satisfactory theory of elections therefore seems to require more structure than that presumed in the benchmark model. Important approaches in this regard are to introduce various informational limitations on the part of candidates and voters, to allow voter abstention and to admit the possibility of policy-motivated candidates.

Candidate Uncertainty and Abstention

Candidates for electoral office clearly do not know the details of every individual's preferences or voting criteria. Adding idiosyncratic non-policy characteristics to voters' decisions (for example, their attitude towards the social background of the candidates) and assuming candidates know at best the distribution from which these characteristics are drawn can induce sufficient continuity in candidates' assessments of how policy positions map into vote shares to admit a general equilibrium existence result in pure strategies. Specifically, let $p_i(u_i(a), u_i(b))$ be the probability that voter i votes for candidate A given platforms $a, b \in X$. Then A's expected plurality, given (a, b) and no abstention, is

$$\Pi(a,b) = 2\sum_{i\in N} p_i(u_i(a),u_i(b)) - n.$$

If we assume that p_i is strictly concave increasing (respectively, convex decreasing) in $u_i(a)$ (respectively, $u_i(b)$) and that u_i is strictly concave, there exists a unique equilibrium in pure candidate strategies and the equilibrium platforms coincide (Enelow and Hinich 1982; Coughlin 1992). More importantly, it turns out that the policy on which both candidates converge in equilibrium maximizes weighted aggregate utility (Couglin and Nitzan 1981; Banks and Duggan 2005).

On the other hand, if candidates $j \in \{A, B\}$ are policy-motivated and seek to maximize their expected utility defined by

$$\begin{aligned} E\left[u_j \middle| a,b\right] &= \Pr[A \text{ wins} \mid a,b]u_j(a) \\ &\quad + \Pr[B \text{ wins} \mid a,b]u_j(b), \end{aligned}$$

then, under suitable regularity conditions on the probabilities of winning, candidate convergence is not assured (Wittman 1977; Calvert 1985). Unfortunately, such regularity conditions are unlikely: for instance, if there is no abstention, then

$$\begin{aligned} &\Pr[A \text{ wins} \mid a,b] \\ &= \sum_{\substack{M\subseteq N \\ |M|>n/2}} \left[\prod_{i\in M} p_i(u_i(a),u_i(b)) \right. \\ &\quad \left. \times \prod_{j\in N/M} \left[1 - P_j\left(u_j(a),u_j(b)\right)\right]\right] \end{aligned}$$

which is not at all nicely behaved.

A conceptual difficulty with the probabilistic voting approach is that it seems ad hoc. Although idiosyncratic components to voters' decision calculus are plausible, the assumptions required on the distributions of such idiosyncrasies to insure existence – that they are uncorrelated with individuals' policy preferences and induce the appropriate concavity properties – are stringent. In particular, if the candidates' uncertainty regards voter policy preferences rather than some non-policy idiosyncrasies, then candidate objective functions again become discontinuous, leading to a breakdown in equilibrium (Ball 1999). An alternative approach in the same spirit is to say nothing about voter idiosyncrasies at the individual level but rather to assume that the winner depends on policy-oriented voting over platforms and the true state of the world, known only up to its distribution at the time of the election (Roemer 2001). The interpretation here is that the realized preference profile of voter preferences is conditional on the state. For example, if those who live closest to voting stations are the most likely to vote in bad weather, then the effective distribution of voter preferences is conditional on the weather. To insure electoral equilibria then requires imposing particular conditions directly on the distribution of states.

Voting is not costless and this fact gives rise to a problem for rational choice theoretic models of participation in large elections: given that the probability of being pivotal is negligible in large elections, the net benefit of voting is negative (Feddersen 2004, provides a recent survey of the literature on turnout). However, assuming that voting costs and (possibly) policy preferences are private information to individuals, with only their joint distribution being common knowledge, induces uncertainty on the part of both voters and candidates sufficient to yield existence of equilibrium in a model with all agents being fully instrumentally rational (Ledyard 1984; Myerson 2000). The idea is to note that, for any fixed and distinct pair of platforms $(0, b) \in X^2$, voters are confronted with a strategic decision whether to abstain or to vote for their favoured candidate. In a voter equilibrium relative to (a, b), the probability of being pivotal induces a level of expected turnout which in turn justifies the probability of being pivotal. Candidates then choose their platforms to maximize their respective expected pluralities, recognizing the implications for turnout from any pair of platforms. In this setting, under various assumptions on the joint distribution of preferences and voting costs, there exists a unique equilibrium when voter preferences are concave and, in equilibrium, both candidates adopt the policy that maximizes aggregate utility and turnout is zero. In effect, this model produces an efficiency result for electoral competition analogous to the First Welfare Theorem for competitive markets.

Voter Uncertainty and Commitment

Although voters are uncertain about the behaviour of other voters in the costly voting model discussed above, they are fully informed about the candidates' platforms and the policy that the winner will implement. Developing an understanding of how voter uncertainties about candidate policies and intentions affect electoral competition in multidimensional policy spaces has proved very difficult. Instead, most of the insights to date derive from one-dimensional models where, under complete information, the median voter theorem applies.

Suppose candidates for office have policy preferences over a one-dimensional policy space just like voters, but that these preferences are unknown to the electorate. Specifically, a candidate's type $t \in R$ parameterizes the candidate's preferences over policies and is private information; voters know only the distribution from which t is drawn. In a two-candidate election, the candidates, knowing their own types, simultaneously choose policy platforms on which to campaign; voters observe the platforms, update their beliefs about the likely types of the candidates, and vote accordingly. The winner is then free to implement any policy she chooses as government policy; in particular, she has an incentive to implement her ideal point. Assume that there is a cost to implementing any policy other than that on which the winner is elected and, further, that this cost increases with the distance between the electoral platform and the implemented policy. Then

Banks (1990) shows that, in any (appropriately refined) sequential equilibrium, relatively extreme candidate types, that is, those with preferences far from those of the median voter, offer revealing platforms but there is pooling on the median voter's most preferred outcome by an interval of relatively moderate types.

The model therefore predicts candidate divergence in equilibrium precisely in competitions in which at least one extremist is running for office; indeed, the observed equilibrium platforms in such instances can be far from the median. More importantly, as the cost of implementing any policy distinct from that on which the election was won goes to zero, the interval of types pooling on the median voter's preferred policy expands to include the entire type space. Conversely, if the cost goes to infinity, the interval of types pooling on the median voter's most preferred policy contracts to the median's ideal point. *Inter alia*, this result highlights the central role of policy commitment in models of electoral competition. If we leave aside concerns with legislative coalition formation and so forth, elected candidates are free to implement any policy they choose once in office. In the benchmark model with full information and purely office-motivated candidates, there is no reason for an elected official to implement anything other than her electoral platform. This is not so if candidates' motivations or preferences are unclear to voters. Unless candidates can make credible commitments to implement their electoral platforms conditional on being elected, electoral platforms are at best signals of candidate intentions; and there is no obvious way for candidates to make such commitments.

If candidates are assumed to have adopted distinct platforms and members of the electorate have private and (possibly) asymmetric information about which candidate is most likely to be their *ex post* preferred candidate, then Feddersen and Pesendorfer (1996, 1997) prove that, as the electorate becomes arbitrarily large, the winning candidate is almost surely the winning candidate under complete information. This is a remarkable result and suggests that questions of commitment and voter uncertainty need not be problematic in large elections. On the other hand, allowing for some strategic platform choice by candidates attenuates the result (Razin 2003; Gul and Pesendorfer 2004).

Dynamics

Parties, if not candidates, are often long-lived, and winning platforms in one election may be empirically hard to change for a following election. Assume the benchmark model of elections above is iterated over an uncountably infinite number sequence of elections $t = 1, 2, \ldots$ with all voters and the two candidates being myopic. Assume in addition that the electoral platform on which the period-t incumbent won office is necessarily the platform on which the incumbent contests election $t + 1$. The opposition candidate's choice of platform in period $t + 1$, however, is unconstrained. Then the non-existence results for equilibria discussed above imply that the two candidates alternate in office over time. More interesting is the fact that the winning platform converges to a neighbourhood of the minmax set, a centrally located set of alternatives that coincides with the majority core when the latter is non-empty: for any alternative x, let $\gamma(x)$ denote the maximal number of votes that any alternative policy y could attract in a pairwise vote against x (equilibrium non-existence implies $\gamma(x) > n/2$ for all $x \in X$); then the minmax set is the set of policies $x \in X$ for which $\gamma(x)$ is minimal (Kramer 1977).

The myopia assumption and the constraint on an incumbent's policy choice underlying the preceding result are not very satisfactory. A strategically richer framework is proposed by Banks and Duggan (2002). In their set-up, all individuals are far-sighted and instrumentally rational. Individual preferences are private information up to the common knowledge that utilities are continuous and strictly concave over the multidimensional policy space. The $t = 1$ incumbent is chosen randomly and implements a policy $x_1 \in X$; the incumbent's name and the policy choice are observed by all voters. In the period $t = 2$ election, the incumbent faces a randomly chosen challenger whose name is observed by voters. Voters then vote for one or other of the competitors. The plurality winner becomes the incumbent and (freely) implements a policy x_2

$\in X$; there is no restriction on the $t = 1$ incumbent's choice should he or she win a second term in office. This process then repeats for $t > 2$. The authors prove the existence of (stationary subgame perfect) equilibria in which voters employ a simple cut-off rule, that is, vote for the incumbent if her previous policy choice at least achieves an endogenously determined reservation utility value. The main result here is that there is eventual policy persistence in that the distribution of policy outcomes converges to a fixed platform as t goes to infinity. However, while this platform is necessarily centrally located because it must be acceptable to a majority of the population, there is no assurance in the multidimensional policy space that it is uniquely defined.

Multiple Candidate Competition

Fixed Number of Candidates

Political competitions with exactly two candidates are relatively unusual. In general there are at least two candidates in any election, and the possibility of multiple electoral competitors raises a variety of issues that are largely irrelevant when considering elections with two given candidates seeking a single office. For example, questions of candidate participation, or the number of candidates, in an election are finessed by assuming a given two-candidate contest, and proportional representation schemes for determining electoral success are irrelevant when only one office is at stake. As a result, important questions about the relative merits of various electoral rules cannot be addressed. Similarly, if the election is for a legislature and legislative policy decisions require, as is typical, majority support of the elected legislators, then rational voters and candidates make their respective electoral decisions taking account of the subsequent legislative bargaining and committee decision-making (Austen-Smith and Banks 1988; Baron and Diermeier 2001). Addressing these issues, among others, requires admitting a more general class of electoral rule for multi-candidate competition and providing a more complex analysis of voter behaviour.

There is fixed set of candidates $M = \{1, \ldots, m\}$ who compete for $l \geq 1$ elected offices, $m < l$, in an election decided by a normalized rank scoring rule. A normalized rank scoring rule for a fixed number of candidates m is defined by a vector $s = (s_1, \ldots, sm)$ such that $1 = s_1 \geq s_2 \geq \ldots \geq s_{m-1} \geq s_m = 0$ and a mapping that assigns a set of $l \geq 1$ winners for any profile of permissible ballots, where a permissible ballot is any permutation of the vector $(1, s_2, \ldots, s_{m-1}, 0)$. The normalization here refers to the joint restrictions $s_1 = 1$ and $s_2 = 0$ and is purely a convenience; the defining characteristics of rank scoring rules are that $s_t \geq s_{t+1}$ for all $t = 1, \ldots, m-1$ and $s_1 > s_m$. Not all rules of interest are rank scoring rules. For instance, approval voting is a scoring rule but not a rank scoring rule; under approval voting, the restriction that $s_1 > s_m$ is not required so voters may vote for, or approve of, any and all candidates should they so choose. Examples of rank scoring rules include single non-transferable voting ($s_t = 0$, all $t > 1$), a generalization of the simple plurality rule (where $l = 1$); single negative voting ($s_t = 1$, all $t < m$); and the Borda rule ($s_t = [m - t]/[m - 1]$, all $t = 2, \ldots, m - 1$). In each of these examples, the l top-scoring candidates are the winners, with ties being broken randomly.

An individual votes sincerely under a scoring rule if she always assigns higher scores to more preferred candidates. If we assume a one-dimensional policy space, say $X = [0,1]$, sincere voting and that candidates maximize expected plurality, now defined to be the difference between their vote share (aggregate score) and the maximum vote share among their competitors, Cox (1987, 1990) establishes a connection between the existence of equilibria in which all candidates converge and the average score $C(s; m) = \sum_{j=1}^{j=m} s_j/m$, the Cox threshold (Myerson 1999).

Specifically, suppose there is a continuum of voters with interior most-preferred policies (or ideal points) distributed over a policy space according to a strictly increasing and differentiable cdf, F and let α_t be the quantile implicitly defined by $\int_0^{\alpha_t} dF(x) = t, t \in [0, 1]$; then there exists an equilibrium in which all candidates adopt the platform αt if and only if $[1 - C(s;\, m)] \leq t \leq C(s;\, m)$.

The Cox threshold for single non-transferable voting is $1/m$, it is 1/2 for the Borda rule and it is $(m - 1)/m$ for single negative voting. The median voter theorem is therefore a direct implication of the theorem. Moreover, writing the Cox threshold equivalently as $C(s; m) \equiv \bar{s} = 1 - (s_1 - \bar{s})/(s_1 - s_m)$ makes clear that the threshold is decreasing in the extent to which the scoring rule provides an incentive to be ranked first rather than average relative to being ranked first rather than last. Hence, Cox's result implies that, the greater this incentive is, the greater is the incentive for candidates to differentiate their platforms.

While sincere voting is the unique undominated strategy for individuals when there are only two candidates, it is not obviously rational when there are the more than two candidates. Indeed, the fact that people typically try to avoid 'wasting' their vote by voting for an almost sure loser suggests that strategically rational voting is substantively significant. An important observation in this respect generalizes Duverger's Law, namely, that single non-transferable voting for a single office promotes only two-candidate competition. Assuming there are $m > 2$ candidates with fixed and distinct platforms competing for $l < m$ offices, Cox (1994) proves that voting equilibria in undominated strategies have the following form: the top l candidates receive identical (strictly positive) vote shares, with all other candidates receiving either no votes or the same vote share as the candidate with the $(l + 1)^{th}$ highest vote share, which in turn is less than or equal to that of the candidates with the l highest vote shares. Moreover, although not proved formally, equilibria in which the vote-share of the $(l + t)^{th}$ most successful candidate, $t > 1$, is positive are almost surely not robust, as any shock will move votes away from this candidate to one of the $l + 1$ top-ranked competitors.

Candidate Entry

The last result on Duverger's Law raises a question as to why candidates who are almost surely going to lose enter the electoral competition at all. In view of the canonical model of electoral competition, one natural starting point for understanding candidate entry is to fix a set of potential candidates, assume each candidate is concerned only with winning office, and ask what platforms such candidates would adopt if entry is costly and voters are strategically rational rather presumed than to vote sincerely under all circumstances. Then purely office-motivated candidates have no incentive to enter an electoral contest if they are sure to lose. On the assumption that voters have strictly concave preferences over a one-dimensional issue space, it can be shown that the number of entrants is constrained only by the ratio of benefits from holding office to the costs of entry and that all entrants adopt the median voter's most preferred policy as their platform (Feddersen et al. 1990). This result is in stark contrast to the implications discussed earlier of assuming a fixed number of competing candidates and sincere voting. As such, the value of this benchmark entry model derives from its appeal less as a reflection of any empirical reality (which is limited at best) and more as an analytical robustness check on models of multi-candidate elections that presume fixed numbers of candidates and sincere voting. While the latter certainly illuminate some incentives facing strategic candidates with several competitors, equilibrium predictions supported by the models appear fragile.

An alternative, and more plausible, approach to presuming a fixed pool of potential candidates is to recognize that candidates are voters too and to suppose the pool is exactly the electorate itself: every individual voter is eligible to run for office should he or she so choose (Osborne and Slivinski 1996; Besley and Coate 1997). In such a model candidate preferences derive directly from individual policy preferences, implying that the entry decision and the decision on which platform to run for office conditional on entering an election are equivalent. At least in the static setting without commitment, citizens who enter the race implement their respective ideal policies should they win office; it immediately follows that policy convergence is impossible if multiple citizens declare a candidacy, unless all such individuals share the identical ideal point. An equilibrium in this framework is a mutually consistent list of best-response decisions for individuals regarding whether or not to enter the election and, conditional on the realized slate of entrants, on how to vote. Given complete information on individuals' preferences and the inability of entrants to commit to implement any policy other than their ideal points, establishing a fairly general existence theorem for multidimensional policy spaces is straightforward. Furthermore, since candidates are also citizens with policy preferences, there are circumstances under which it is more rational for an individual to contest a costly election than when individuals are concerned only with winning office. Specifically, despite being sure that he or she will not win, an individual might enter an electoral race to affect the expected policy outcome by implicitly blocking the entry or exit of other potential or declared candidates.

The citizen-candidate perspective on political competition is quite intuitive and simultaneously gives rise to a theory of candidate entry and a theory of candidate preferences. On the other hand, the implication that, other things being equal, declared candidates are locked into implementing their ideal points should they win is restrictive. As a matter of fact, candidates do credibly adjust their policy positions and, even should they not do so, a theory of platform selection predicated exclusively on an individual's exogenously given policy preferences effectively reduces any account of collective policy outcomes to an account of which particular individuals choose to run for office. Although understanding who chooses to seek election is clearly important to a full theory of electoral politics, an account of collective choice based entirely on such a foundation is less compelling.

A different approach within the spirit of the citizen-candidate theory is to eschew candidates and parties altogether. Instead, the set of 'potential candidates' is assumed to be the full set of feasible policy outcomes, with each individual permitted to vote for any one policy or not at all. Assuming away candidates as agents or, equivalently, assuming there is a candidate for every possible alternative in the policy space, seems, at least prima facie, to be unreasonable for any sort of theory of electoral competition with a large number of voters. In so far as the focus of a model of elections is on understanding the behaviour of particular agents with given objectives (for example, maximizing the probability of winning), or on understanding the behavior of historically established political parties in a constrained (for example, two-party) environment, such a presumption is justified. However, if the focus is on understanding the deeper implications of using a given voting scheme to determine collective policy choices, then dispensing at the outset with the intermediary steps involved in candidates choosing platforms is reasonable. And in this respect the implications of plurality rule under costly voting and abstention are subtle and striking: in every voting equilibrium with risk-averse (and strategically rational) voters and a multidimensional policy space, exactly two divergent policy positions receive any votes at all (Feddersen 1992). Despite the absence of candidates or parties, strategic and costly voting under plurality rule leads to equilibria with two distinct platforms and positive turnout. The precise location of any given pair of equilibrium platforms, however, is in general indeterminate.

Concluding Comments

Despite the large literature on the spatial theory of elections, our understanding of political competition is still relatively primitive in many respects. There is, for example, much to be learned from dynamic models, and a compelling theory of candidate entry has yet to be developed. Similarly, a tractable theory of strategic voting in large populations, an essential component of a satisfactory theory of political competition, remains elusive.

See Also

- ▶ Collective Rationality
- ▶ Political Institutions, Economic Approaches to

Bibliography

Austen-Smith, D., and J. Banks. 1988. Elections, coalitions and legislative outcomes. *American Political Science Review* 82: 405–422.

Ball, R. 1999. Discontinuity and nonexistence of equilibrium in the probabilistic spatial voting model. *Social Choice and Welfare* 16: 533–556.

Banks, J. 1990. A model of electoral competition with incomplete information. *Journal of Economic Theory* 50: 309–325.

Banks, J., and J. Duggan. 2002. A multidimensional model of repeated elections. Working paper. Department of Political Science, University of Rochester.

Banks, J., and J. Duggan. 2005. Probabilistic voting in the spatial model of elections: The theory of office-motivated candidates. In *Social choice and strategic decisions: Essays in Honor of Jeffrey S. Banks*, ed. D. Austen-Smith and J. Duggan. Berlin: Springer.

Banks, J., J. Duggan, and M. Le Breton. 2002. Bounds for mixed strategy equilibria and the spatial model of elections. *Journal of Economic Theory* 103: 88–105.

Baron, D., and D. Diermeier. 2001. Elections, governments and parliaments in proportional representation systems. *Quarterly Journal of Economics* 116: 933–967.

Besley, T., and S. Coate. 1997. An economic model of representative democracy. *Quarterly Journal of Economics* 112: 85–114.

Black, D. 1958. *The theory of committees and elections*. Cambridge: Cambridge University Press.

Calvert, R. 1985. Robustness of the multidimensional voting model: Candidate motivations, uncertainty and convergence. *American Journal of Political Science* 29: 69–95.

Coughlin, P. 1992. *Probabilistic voting theory*. Cambridge: Cambridge University Press.

Couglin, P., and S. Nitzan. 1981. Electoral outcomes with probabilistic voting and Nash social welfare maxima. *Journal of Public Economics* 15: 113–121.

Cox, G. 1987. Electoral equilibrium under alternative voting institutions. *American Journal of Political Science* 31: 82–108.

Cox, G. 1990. Centripetal and centrifugal incentives in electoral systems. *American Journal of Political Science* 34: 903–935.

Cox, G. 1994. Strategic voting equilibria under the single nontransferable vote. *American Political Science Review* 88: 608–621.

Davis, O., and M. Hinich. 1966. A mathematical model of policy formation in a democratic society. In *Mathematical applications in political science II*, ed. J. Bernd. Dallas: Southern Methodist University Press.

Davis, O., and M. Hinich. 1967. Some results related to a mathematical model of policy formation in a democratic society. In *Mathematical applications in political science III*, ed. J. Bernd. Dallas: Southern Methodist University Press.

Downs, A. 1957. *An economic theory of democracy*. New York: Harper.

Duggan, J., and M. Fey. 2005. Electoral competition with policy-motivated candidates. *Games and Economic Behavior* 51: 490–522.

Duggan, J., and M. Jackson. 2004. Mixed strategy equilibrium and deep covering in multidimensional electoral competition. Working Paper. Department of Political Science, University of Rochester.

Enelow, J., and M. Hinich. 1982. Nonspatial candidate characteristics and electoral competition. *Journal of Politics* 44: 115–130.

Feddersen, T. 1992. A voting model implying Duverger's Law and positive turnout. *American Journal of Political Science* 36: 938–962.

Feddersen, T. 2004. Rational choice theory and the paradox of not voting: A review. *Journal of Economic Perspectives* 18: 99–112.

Feddersen, T., and W. Pesendorfer. 1996. The swing voter's curse. *American Economic Review* 86: 408–424.

Feddersen, T., and W. Pesendorfer. 1997. Voting behavior and information aggregation in elections with private information. *Econometrica* 65: 1029–1058.

Feddersen, T., I. Sened, and S. Wright. 1990. Rational voting and candidate entry under plurality rule. *American Journal of Political Science* 34: 1005–1016.

Gul, F., and W. Pesendorfer. 2004. Partisan politics and aggregation failure with ignorant voters. Working paper. Princeton: Department of Economics, Princeton University.

Hotelling, H. 1929. Stability in competition. *Economic Journal* 39: 41–57.

Kramer, G. 1977. A dynamical model of political equilibrium. *Journal of Economic Theory* 16: 310–334.

Laffond, G., J.-F. Laslier, and M. Le Breton. 1993. The bipartisan set of a tournament game. *Games and Economic Behavior* 5: 182–201.

Ledyard, J. 1984. The pure theory of large two-candidate elections. *Public Choice* 44: 7–43.

McKelvey, R. 1979. General conditions for global intransitivities in formal voting models. *Econometrica* 47: 1086–1112.

McKelvey, R. 1986. Covering, dominance and institution-free properties of social choice. *American Journal of Political Science* 30: 283–314.

Miller, N. 1980. A new solution set for tournaments and majority voting. *American Journal of Political Science* 24: 68–96.

Myerson, R. 1999. Theoretical comparisons of electoral systems. *European Economic Review* 43: 671–697.

Myerson, R. 2000. Large Poisson games. *Journal of Economic Theory* 94: 7–45.

Myerson, R. 2002. Comparison of scoring rules in Poisson voting games. *Journal of Economic Theory* 103: 217–251.

Osborne, M., and A. Slivinski. 1996. A model of political competition with citizencandidates. *Quarterly Journal of Economics* 111: 65–96.

Plott, C. 1967. A notion of equilibrium and its possibility under majority rule. *American Economic Review* 57: 787–806.

Razin, R. 2003. Signaling and election motivations in voting model with common values and responsive candidates. *Econometrica* 71: 1083–1120.

Roemer, J. 2001. *Political competition: Theory and applications*. Cambridge: Harvard University Press.

Schofield, N. 1983. Generic instability of majority rule. *Review of Economic Studies* 50: 695–705.
Wittman, D. 1977. Candidates with policy preferences: A dynamic model. *Journal of Economic Theory* 14: 180–189.
Wittman, D. 1983. Candidate motivation: A synthesis of alternatives. *American Political Science Review* 77: 142–157.

'Political Economy'

Peter Groenewegen

Keywords

Baran, P.A.; Bentham, J.; Cairnes, J. E.; Classical economics; Dobb, M.H.; Economics; Engels, F.; Hearn, W. E.; Jevons, W.S.; MacLeod, H.D.; Marshall, A.; Marshall, M.P.; Marx, K. H.; McCulloch, J. R.; Mill, J.; Mill, J. S.; Mirabeau, V. R., Marquis de; Normative economics; Physiocracy; Political arithmetic; Political economy; Quesnay, F.; Robbins, L. C.; Robinson, J.V.; Senior, N. W.; Steuart, J.; Surplus approach; Torrens, R.

JEL Classifications

B0

This article provides a survey of the origin of the term 'political economy' and its changes in meaning, emphasizing in particular its first modern usage in the 18th century, its demise from the end of the 19th century, when it was gradually replaced by the word 'economics', and its revival in a variety of forms, largely during the 1960s, which have altered its meaning from more traditional usage. What follows is therefore largely definitional and etymological, designed to indicate the lack of precise meaning associated with both the term, 'political economy' and its more modern synonym, 'economics'.

The origin of words starting with 'econom' is Greek, from *eco* meaning 'house' and *nom* meaning 'law' in the sense appropriate to astronomy when it deals with 'the law and order of the stars' (Cannan 1929, p. 37). The traditional meaning of *oikonomike* or economics, was therefore 'household management'. Aristotle (1962, p. 30) used it in this sense when analysing households as 'three pairs: master and slave, husband and wife, father and children'. This meaning persisted in moral philosophy until the middle of the 18th century, for example, in Hutcheson (1755) and Smith (1763, p. 141). The Latin *oeconomia* likewise meant management of household affairs, extended to management in general including orderly arrangement of speech and composition. The French *oeconomie* or *économie* took over this wider meaning of management from the Latin and when combined with *politique* it signified public administration or management of the affairs of state. Arthur Young (1770) applied this wider meaning in the title of a treatise on agricultural management. Using 'economy' as a synonym for 'thrift', 'frugality' and careful management of the finances of households and other organizations also derives from the Latin adaptation. 17th-century concern with nation building gave the term) 'public administration' a wider scope, and given developments in France under Henry IV and Richelieu it is not surprising that the term 'political economy' made its first appearance there. This first use is generally attributed to Montch'retien (1615), but King (1948) indicates prior use in Mayerne-Turquet (1611). Because the relationship between state and economy it signified was so appropriate to the times, King suggests that other, perhaps earlier, uses may be found. Petty (1691, p. 181; cf. 1683, p. 483) used the term in England. As Cannan (1929, p. 39) surmised, he could as well have used 'political economy' as 'political anatomy' to describe his analysis of the Irish economy, considering he used 'political arithmetick' for the art of making more precise statements on the political economy of nations, interpreted as their comparative strengths (cf. Verri 1763, pp. 9–10, who speaks of the science of political economy in this manner). Cantillon (1755, p. 46) referred to an 'oeconomy' in the sense of an economic organism in which classes exist as interdependent units, but his book remained an 'Essay on Commerce'.

More precise formulations of political economy as a science of economic organization, though with continuing connotations of management, regulation and even orderly natural laws, are found in Physiocracy. Quesnay's early usage generally implies the traditional meanings, but in addition he applied the term to include discussions of the nature of wealth, its reproduction and distribution. This double meaning is particularly evident in his *Tableau économique*. It is therefore no accident that Mirabeau (1760) spoke of *économie politique* 'as if it consisted of a dissertation on agriculture and public administration as well as on the nature of wealth and the means of procuring it' (Cannan 1929, p. 40). During the subsequent decades the second meaning became more dominant, the word 'science' was added to it (an innovation attributed to Verri 1763, p. 9) and by the 1770s it almost exclusively referred to the production and distribution of wealth in the context of management of the nation's resources.

Sir James Steuart (1767) is the first English economist to put 'political economy' into the title of a book. Its introductory chapter explained that just as 'Oeconomy in general, is the art of providing for all the wants of the family', so the science of political economy seeks 'to secure a certain fund of subsistence for all the inhabitants, to obviate every circumstance which may render it precarious; to provide every thing necessary for supplying the wants of the society, and to employ the inhabitants . . . in such a manner as naturally to create reciprocal relations and dependencies between them, so as to make their several interests lead them to supply one another with reciprocal wants' (1767, pp. 15, 17). Steuart's full title gave the subject matter to be covered: 'population, agriculture, trade, industry, money, coin, interest, circulation, banks, exchange, public credit and taxes'. In 1771 Verri published *Reflections on Political Economy*, the preface of which referred to a new department of knowledge called political economy. Although Smith did not use 'political economy' in his title the introduction and plan of his book refers to 'different theories of political economy' and at the start of Book IV he defined the term as 'a branch of the science of a statesman or legislator' with the twofold objectives of providing 'a plentiful revenue or subsistence for the people . . . [and] to supply the state or commonwealth with a revenue sufficient for the publick services' (Smith 1776, pp. 11, 428). Elsewhere (1776, pp. 678–9) Smith indicated that he saw political economy as an inquiry into the nature and causes of the wealth of nations or, as the Physiocrats had initially suggested, the science of the nature, reproduction, distribution and disposal of wealth.

The association of the science, political economy, with material welfare proved to be particularly hardy, as was its association with the art of legislation. Bentham (1793–5, p. 223) put the matter concisely when he argued, 'Political Economy may be considered as a science or as an Art. But in this instance as in others, it is only as a guide to the art that the science is of use'. Torrens (1819, p. 453) also called it 'one of the most important and useful branches of science' while James Mill (1820, p. 211) and McCulloch (1825, p. 9) defined it as a systematic inquiry into the laws regulating the production, distribution, consumption and exchange of commodities or the products of labour. 'Confounding' the art with the science was criticized by Senior (1836, p. 3) as being detrimental to its development, a position likewise taken by John Stuart Mill (1831–3) and which also reaffirmed its moral and social nature. In this influential essay, Mill (1831–3, p. 140) defined political economy as 'the science which traces the laws of such of the phenomena of society as arise from the combined operations of mankind for the production of wealth, in so far as those phenomena are not modified by the pursuit of any other object'. This position was more or less adhered to in his later *Principles* (1848, p. 21), when he defined its subject matter as 'the laws of Production and Distribution, and some of the practical consequences deducible from them . . . '. Cairnes (1875, p. 35) condensed this to the statement that 'Political Economy . . . expounds the laws of the phenomena of wealth.'

The middle of the 19th century saw two criticisms of this meaning of political economy. Marx (1859, p. 20) identified the study of political economy with a search for 'the anatomy of civil society' or, as Engels (1859, p. 218) put in his review of this book, 'the theoretical analysis of modern bourgeois society'. This preserved the name but criticized the scope and method of political economy. Others suggested the name be changed because it had become misleading. Hearn (1863) put forward *Plutology* or the theory of efforts to satisfy human wants; MacLeod (1875) proposed 'economics', defining it as the 'science which treats of the laws which govern the relations of exchangeable quantities', a nomenclature of whose virtues he successfully persuaded Jevons (Black 1977, p. 115). When in 1879 the Marshalls published an elementary political economy text, they called it *The Economics of Industry*. The new name of MacLeod and the Marshalls was favourably referred to in the second edition of Jevons's *Theory* (1879, p. xiv) because of convenience and scientific nicety (it matched mathematics, ethics and aesthetics) and Jevons's last published book (Jevons 1905) bore the title *Principles of Economics*. Although Cannan (1929, p. 44) claimed Marshall (1890) induced acceptance of the new name, this only came with the later editions, and the change was not completed until the early 1920s (Groenewegen 1985). Even then, Marshall (1890, p. 1) appeared to treat the two names as synonyms: 'Political Economy or Economics is a study of mankind in the ordinary business of life; it examines that part of individual and social action which is most closely connected with the attainment and with the use of the material requisites of well-being.'

Just as J.S. Mill (1831–3, pp. 120–1) had attempted retrospective codification of scope and method in the 1820s, so Robbins (1932, p. 16) redefined economics in its marginalist form as 'the science which studies human behaviour as a relationship between ends and scarce means which have alternative uses'. This did more than supply a meaning for the new term, 'economics'. It destroyed the view classical economists had of their science, as Myint (1948) clearly pointed out. Others (for example, Knight 1951, p. 6) complained that Robbins's definition neglected the link between economics and the 'individualistic or "liberal" outlook on life, of which "capitalism", or the competitive system, or free business enterprise, is the expression upon the economic side, as democracy on the political'. However, the major drawback of the Robbins definition was its irreconcilability with Keynes's work with its proof of the possibility of unemployment equilibrium and hence contradicting Robbins's requirement for the existence of an economic problem that resources have to be scarce. Modern mainstream definitions of economics (Rees 1968; Samuelson 1955, p. 5) have simply combined the Robbinsian resource allocation problem with the new economics of employment, inflation and growth developed from Keynes's work.

Robbins's definition also aimed to make economics a 'system of theoretical and positive knowledge' (Fraser 1937, p. 30), preferring to reserve the older name, 'political economy', for applied topics such as monopoly, protection, planning and government fiscal policy, subjects included in his essays on political economy (Robbins 1939). Although Schumpeter (1954) held a similar opinion he was careful to warn that 'political economy meant different things to different writers, and in some cases it meant what is now known as economic theory or "pure" economics' (p. 22). These views of political economy conflict with the pragmatic Cambridge outlook on economics, derived from Marshall's description of economics as 'an engine for the discovery of concrete truth', encapsulated by Keynes (1921, p. v) in his famous introduction to the Cambridge Economics Handbooks: 'Economics is a method rather than a doctrine, an apparatus of the mind, a technique of thinking which helps its possessor to draw correct conclusions.' This sentiment is concisely summarized by Joan Robinson's view of economics (1933, p. 1) as 'a box of tools'.

Marxists had never abandoned the older terminology of political economy. Dobb (1937, p. vii) defended 'political economy' against the new term 'economics' because its controversies 'have meaning as answers to certain questions of an essentially practical kind', associated with the 'nature and behaviour' of the capitalist system. Likewise, Baran (1957, p. 131) argued for a 'political economy of growth' because an 'understanding of the factors responsible for the size and the mode of utilization of the social surplus . . . [is] a problem, not even approached in the realm of pure economics'. For the classical economists, use of the surplus had been a major research question. Political economy is therefore a very appropriate title for the endeavours of some contemporary economists to resurrect both practical and theoretical aspects of the classical tradition in what they describe as the surplus approach.

By the 1960s the radical libertarian right from Chicago and the Center for the Study of Public Choice at Virginia Polytechnic appeared to have appropriated the title 'political economy' for their wide application of Robbins's (1932) injunction that analysis in terms of '*alternatives*' is the key distinguishing feature of economics. This effectively replaced Robbins's question 'what is or is not economic in nature' with the far wider one of 'what can economics contribute to our understanding of this or that problem?' This opens up the way for an economics of 'family life, child rearing, dying, sex, crime, politics and many other topics' which some of its practitioners identify with Adam Smith's research agenda (McKenzie and Tullock 1975, p. 3). Others continue to associate the term 'with the specific advice given by one or more economists . . . to governments or to the public at large either on broad policy issues or on particular proposals' or, alternatively, as another term for 'normative economics' (Mishan 1982, p. 13).

At the approach of the 21st century, both terms – 'political economy' and 'economics' – survive. During their existence, both have experienced changes of meaning. Nevertheless, they can still essentially be regarded as synonyms, a feature of this nomenclature reflecting an interesting characteristic of the science it describes. In its sometimes discontinuous development, economics or political economy has invariably experienced difficulties in discarding earlier views, and traces of old doctrine are intermingled with the latest developments in the science.

Bibliography

Aristotle. 1962. *The politics*. Trans. J.E. Sinclair. Harmondsworth: Penguin Classics.

Baran, P.A. 1957. *The political economy of growth*. Harmondsworth: Penguin Books, 1973.

Bentham, J. 1793–5. Manual of political economy. In *Jeremy Bentham's economic writings*, ed. W. Stark. London: George Allen & Unwin, 1952.

Black, R.D.C., ed. 1977. *Papers and correspondence of William Stanley Jevons: Correspondence 1873–78*. London: Macmillan for the Royal Economic Society.

Cairnes, J.E. 1875. *The character and logical method of political economy.* London. Reprinted, New York: Kelly, 1965.

Cannan, E. 1929. *A review of economic theory.* London: P.S. King & Son.

Cantillon, R. 1755. In *Essay on the nature of commerce in general.* Ed. H. Higgs. London: Macmillan, 1931.

Dobb, M.H. 1937. *Political economy and capitalism*. London: Routledge.

Engels, F. 1859. Karl Marx's 'A Contribution to the Critique of Political Economy'. *Das Volk*, No. 14, Berlin, August. In Marx (1859).

Fraser, L.M. 1937. *Economic thought and language*. London: A. & C. Black.

Groenewegen, P.D. 1985. Professor Arndt on political economy: A comment. *Economic Record* 61: 744–751.

Hearn, W.E. 1863. *Plutology*. Melbourne: Robertson.

Hutcheson, F. 1755. *A system of moral philosophy.* Glasgow: Robert and Andrew Foulis.

Jevons, W.S. 1879. *The theory of political economy*, 2nd edn. London: Macmillan; Preface in 4th edn, London, 1910.

Jevons, W.S. 1905. *Principles of economics*. London: Macmillan.

Keynes, J.M. 1921. Introduction to Cambridge economic handbooks. In *Money*, ed. D.H. Robertson. London/Cambridge: Cambridge Economic Handbooks.

King, J.E. 1948. The origin of the term 'political economy'. *Journal of Modern History* 20: 230–231.

Knight, F.H. 1951. Economics. In *On the history and method of economics*, ed. F.H. Knight. Chicago: University of Chicago Press, 1963.

MacLeod, H.D. 1875. What is political economy? *Contemporary Review* 25: 871–893.

Marshall, A. 1890. *Principles of economics*. 9th variorum edn, ed. C.W. Guillebaud. London: Macmillan, 1961.

Marshall, A., and M.P. Marshall. 1879. *The economics of industry.* London: Macmillan.

Marx, K. 1859. *A contribution to the critique of political economy.* Introduction by M. Dobb. London: Lawrence & Wishart, 1971.

Mayerne-Turquet, L de. 1611. La Monarchie Aristodémocratique; ou le Gouvernement composé et meslé des trois formes de légitimes républiques. Paris.

McCulloch, J.R. 1825. *Principles of political economy with sketch of the rise and progress of the science*. London: Murray, 1870.

McKenzie, R.B., and G. Tullock. 1975. *The new world of economics: Explorations into the human experience*. Homewood: Irwin.

Mill, J. 1820. *Elements of political economy*, 3rd edn. London, 1926. Reprinted in James Mill, *Selected writings*, ed. D. Wisen. Edinburgh: Oliver & Boyd for the Scottish Economic Society, 1966.

Mill, J.S. 1831–3. On the definition of political economy; and on the method of investigation proper to it. Essay V in J.S. Mill, *Essays on some unsettled questions of political economy*. London, LSE Reprint, 1948.

Mill, J.S. 1848. Principles of political economy with some of their applications to social philosophy. In *Collected works of John Stuart Mill*, ed. J.M. Robson. Toronto: University of Toronto Press, 1965.

Mirabeau, V.R., Marquis de. 1758–60. *L'ami des hommes ou traité de la population*. Avignon and Paris.

Mishan, E.J. 1982. *Introduction to political economy*. London: Hutchinson.

Montchŕetien, A. de. 1615. *Traité de l'économie politique*. Ed. Th. Funck-Brentano. Paris: Plon, 1889.

Myint, H.L.A. 1948. *Theories of welfare economics*. London: Longmans, Green & Co.

Petty, Sir W. 1683. Observations upon the Dublin bills of mortality and the state of that city. In *Economic writings of Sir William Petty*, ed. C.H. Hull. Reprinted, New York: Kelley, 1963.

Petty, Sir W. 1691. The political anatomy of Ireland. In *The economic writings of Sir William Petty*, ed. C.H. Hull. New York: Kelley, 1963.

Rees, A. 1968. Economics. In *International encyclopaedia of the social sciences*, ed. D.L. Sills, vol. 4. New York: Macmillan.

Robbins, L.C. 1932. *An essay on the nature and significance of economic science*. London: Macmillan. 2nd edn, 1935.

Robbins, L.C. 1939. *The economic basis of class conflict and other essays in political economy*. London: Macmillan.

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

Samuelson, P.A. 1955. *Economics*, 3rd edn. New York: McGraw-Hill. 7th edn, 1967.

Schumpeter, J.A. 1954. *History of economic analysis*. London: George Allen & Unwin.

Senior, N.W. 1836. *An outline of the science of political economy*. London: Unwin Library of Economics, 1938.

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

Smith, A. 1776. *An inquiry into the nature and causes of the wealth of nations*. Ed. R.H. Campbell, and A.S. Skinner. Oxford: Oxford University Press, 1976.

Steuart, J. 1767. *An inquiry into the principles of political economy*. Ed. A.S. Skinner. Edinburgh/London: Oliver & Boyd for the Scottish Economic Society, 1966.

Torrens, R. 1819. Mr Owen's plans for relieving the national distress. *Edinburgh Review* 32, October, Article XI.

Verri, P. 1763. Memorie storiche sulla economia pubblica dello stato di Milano. In *Scrittori Classici Italiani di Economia Politica*, Parte Moderna, vol. 17. Milan, 1804.

Verri, P. 1771. In *Reflections on political economy*. Trans. B. McGilvray, ed. P. Groenewegen, Reprints of Economic classics, series 2, no. 4. Sydney: University of Sydney, 1986.

Young, A. 1770. *Rural oeconomy, or essays on the practical parts of husbandry*. London.

'Political Economy' and 'Economics'

Peter Groenewegen

This article provides a survey of the origin of the term 'political economy' and its changes in meaning, emphasizing in particular its first modern usage in the 18th century, its demise from the end of the 19th century, when it was gradually replaced by the word) 'economics', and its revival in a variety of forms, largely during the 1960s, which have altered its meaning from more traditional usage. What follows is therefore largely definitional and etymological, designed to indicate the lack of precise meaning associated with both the term, 'political economy' and its more modern synonym, 'economics'.

The origin of words starting with 'econom' is Greek, from *eco* meaning 'house' and *nom* meaning 'law' in the sense appropriate to astronomy when it deals with 'the law and order of the stars' (Cannan 1929, p. 37). The traditional meaning of *oikonomike* or economics, was therefore 'household management'. Aristotle (1962, p. 30) used it in this sense when analysing households as 'three pairs: master and slave, husband and wife, father and children'. This meaning persisted in moral philosophy until the middle of the 18th century, for example, in Hutcheson (1755) and Smith (1763, p. 141). The Latin *oeconomia* likewise meant management of household affairs, extended to management in general including orderly arrangement of speech and composition. The French *oeconomie* or *économie* took over this wider meaning of management from the Latin and when combined with *politique* it signified public administration or management of the affairs of state. Arthur Young (1770) applied this wider meaning in the title of a treatise on agricultural management. Using 'economy' as a synonym for 'thrift', 'frugality' and careful management of the finances of households and other organizations also derives from the Latin adaptation. 17th-century concern with nation building gave the term 'public administration' a wider scope, and given developments in France under Henry IV and Richelieu it is not surprising that the term 'political economy' made its first appearance there. This first use is generally attributed to Montchrétien (1615), but King (1948) indicates prior use in Mayerne-Turquet (1611). Because the relationship between state and economy it signified was so appropriate to the times, King suggests that other, perhaps earlier uses, may be found. Petty (1691, p. 181 and cf. 1683, p. 483) used the term in England. As Cannan (1929, p. 39) surmised, he could as well have used) 'political economy' as 'political anatomy' to describe his analysis of the Irish economy, considering he used 'political arithmetick' for the art of making more precise statements on the political economy of nations, interpreted as their comparative strengths (cf. Verri 1763, pp. 9–10, who speaks of the science of political economy in this manner). Cantillon (1755, p. 46) referred to an 'oeconomy' in the sense of an economic organism in which classes exist as interdependent units, but his book remained an 'Essay on Commerce'.

More precise formulations of political economy as a science of economic organization, though with continuing connotations of management, regulation and even orderly natural laws, are found in Physiocracy. Quesnay's early usage generally implies the traditional meanings, but in addition he applied the term to include discussions of the nature of wealth, its reproduction and distribution. This double meaning is particularly evident in his *Tableau économique*. It is therefore no accident that Mirabeau (1760) spoke of *économie politique* 'as if it consisted of a dissertation on agriculture and public administration as well as on the nature of wealth and the means of procuring it', (Cannan 1929, p. 40). During the subsequent decades the second meaning became more dominant, the word 'science' was added to it (an innovation attributed to Verri 1763, p. 9) and by the 1770s it almost exclusively referred to the production and distribution of wealth in the context of management of the nation's resources.

Sir James Steuart (1767) is the first English economist to put) 'political economy' into the title of a book. Its introductory chapter explained that just as 'Oeconomy in general, is the art of providing for all the wants of the family', so the science of political economy seeks 'to secure a certain fund of subsistence for all the inhabitants, to obviate every circumstance which may render it precarious; to provide every thing necessary for supplying the wants of the society, and to employ the inhabitants . . . in such a manner as naturally to create reciprocal relations and dependencies between them, so as to make their several interests lead them to supply one another with reciprocal wants', (1767, pp. 15, 17). Steuart's full title gave the subject matter to be covered: 'population, agriculture, trade, industry, money, coin, interest, circulation, banks, exchange, public credit and taxes'. In 1771 Verri published *Reflections on Political Economy*, the preface of which referred to a new department of knowledge called political economy. Although Smith did not use 'political economy' in his title the introduction and plan of his book refers to 'different theories of political economy' and at the start of Book IV he defined the term as 'a branch of the science of a statesman or legislator' with the twofold objectives of providing 'a plentiful revenue or subsistence for the people . . . [and] to supply the state or commonwealth with a revenue sufficient for the publick services' (Smith 1776, pp. 11, 428). Elsewhere (1776, pp. 678–9) Smith indicated that he saw political economy as an inquiry into the nature and causes of the wealth of nations or, as the physiocrats had initially suggested, the science of the nature, reproduction, distribution and disposal of wealth.

The association of the science, political economy, with material welfare proved to be particularly hardy, as was its association with the art of legislation. Bentham (1793–5, p. 223) put the matter concisely when he argued, 'Political Economy may be considered as a science or as an Art. But in this instance as in others, it is only as a guide to the art that the science is of use'. Torrens (1819, p. 453) also called it 'one of the most important and useful branches of science' while James Mill (1820, p. 211) and McCulloch (1825, p. 9) defined it as a systematic inquiry into the laws regulating the production, distribution, consumption and exchange of commodities or the products of labour. 'Confounding' the art with the science was criticized by Senior (1836, p. 3) as being detrimental to its development, a position likewise taken by John Stuart Mill (1831–3) and which also reaffirmed its moral and social nature. In this influential essay, Mill (1831–3, p. 140) defined political economy as 'the science which traces the laws of such of the phenomena of society as arise from the combined operations of mankind for the production of wealth, in so far as those phenomena are not modified by the pursuit of any other object'. This position was more or less adhered to in his later *Principles* (1848, p. 21), when he defined its subject matter as 'the laws of Production and Distribution, and some of the practical consequences deducible from them . . .'. Cairnes (1875, p. 35) condensed this to the statement that 'Political Economy . . . expounds the laws of the phenomena of wealth.'

The middle of the 19th century saw two criticisms of this meaning of political economy. Marx (1859, p. 20) identified the study of political economy with a search for 'the anatomy of civil society' or, as Engels (1859, p. 218) put in in his review of this book, 'the theoretical analysis of modern bourgeois society'. This preserved the name but criticized the scope and method of political economy. Others suggested the name be changed because it had become misleading. Hearn (1863) put forward *Plutology* or the theory of efforts to satisfy human wants; MacLeod (1875) proposed 'economics', defining it as the 'science which treats of the laws which govern the relations of exchangeable quantities', a nomenclature of whose virtues he successfully persuaded Jevons (Black 1977, p. 115). When in 1879 the Marshalls published an elementary political economy text, they called it *The Economics of Industry*. The new name of MacLeod and the Marshalls was favourably referred to in the second edition of Jevons's *Theory* (1879, p. xiv) because of convenience and scientific nicety (it matched mathematics, ethics and aesthetics) and Jevons's last published book (Jevons 1905) bore the title *Principles of Economics*. Although Cannan (1929, p. 44) claimed Marshall (1890) induced acceptance of the new name, this only came with the later editions, and the change was not completed until the early 1920s (Groenewegen 1985). Even then, Marshall (1890, p. 1) appeared to treat the two names as synonyms: 'Political Economy or Economics is a study of mankind in the ordinary business of life; it examines that part of individual and social action which is most closely connected with the attainment and with the use of the material requisites of well-being.'

Just as J.S. Mill (1831–3, pp. 120–1) had attempted retrospective codification of scope and method in the 1820s, so Robbins (1932, p. 16) redefined economics in its marginalist form as 'the science which studies human behaviour as a relationship between ends and scarce means which have alternative uses'. This did more than supply a meaning for the new term, 'economics'. It destroyed the view classical economists had of their science, as Myint (1948) clearly pointed out. Others (e.g. Knight 1951, p. 6) complained that Robbins's definition neglected the link between economics and the 'individualistic or "liberal" outlook on life, of which "capitalism", or the competitive system, or free business enterprise, is the expression upon the economic side, as democracy on the political'. However, the major drawback of the Robbins definition was its irreconcilability with Keynes's work with its proof of the possibility of unemployment equilibrium and hence contradicting Robbins's requirement for the existence of an economic problem that resources have to be scarce. Modern mainstream definitions of economics (Rees 1968; Samuelson 1955, p. 5) have simply combined the Robbinsian resource allocation problem with the new economics of employment, inflation and growth developed from Keynes's work.

Robbins's definition also aimed to make economics a 'system of theoretical and positive knowledge' (Fraser 1937, p. 30), preferring to reserve the older name, 'political economy' for applied topics such as monopoly, protection, planning and government fiscal policy, subjects included in his essays on political economy (Robbins 1939). Although Schumpeter (1954) held a similar opinion he was careful to warn that) 'political economy meant different things to different writers, and in some cases it meant what is now known as economic theory or "pure" economics' (p. 22). These views of political economy conflict with the pragmatic Cambridge outlook on economics, derived from Marshall's description of economics as 'an engine for the discovery of concrete truth', encapsulated by Keynes (1921, p. v) in his famous introduction to the Cambridge Economics Handbooks: 'Economics is a method rather than a doctrine, an apparatus of the mind, a technique of thinking which helps its possessor to draw correct conclusions.' This sentiment is concisely summarized by Joan Robinson's view of economics (1933, p. 1) as 'a box of tools'.

Marxists had never abandoned the older terminology of political economy. Dobb (1937, p. vii) defended 'political economy' against the new term 'economics' because its controversies 'have meaning as answers to certain questions of an essentially practical kind', associated with the 'nature and behaviour' of the capitalist system. Likewise, Baran (1957, p. 131) argued for a 'political economy of growth' because an) 'understanding of the factors responsible for the size and the mode of utilization of the social surplus . . . [is] a problem, not even approached in the realm of pure economics'. For the classical economists, use of the surplus had been a major research question. Political economy is therefore a very appropriate title for the endeavours of some contemporary economists to resurrect both practical and theoretical aspects of the classical tradition in what they describe as the surplus approach.

By the 1960s the radical libertarian right from Chicago and the Center for the Study of Public Choice at Virginia Polytechnic appears to have appropriated the title 'political economy' for their wide application of Robbins's (1932) injunction that analysis in terms of *'alternatives'* is the key distinguishing feature of economics. This effectively replaced Robbins's question 'what is or is not economic in nature' with the far wider one of 'what can economics contribute to our understanding of this or that problem?' This opens up the way for an economics of 'family life, child rearing, dying, sex, crime, politics and many other topics' which some of its practitioners identify with Adam Smith's research agenda (McKenzie and Tullock 1975, p. 3). Others continue to associate the term 'with the specific advice given by one or more economists . . . to governments or to the public at large either on broad policy issues or on particular proposals' or, alternatively, as another term for 'normative economics' (Mishan 1982, p. 13).

At the approach of the 21st century, both terms – 'political economy' and 'economics' – survive. During their existence, both have experienced changes of meaning. Nevertheless, they can still essentially be regarded as synonyms, a feature of this nomenclature reflecting an interesting characteristic of the science it describes. In its sometimes discontinuous development, economics or political economy has invariably experienced difficulties in discarding earlier views, and traces of old doctrine are intermingled with the latest developments in the science.

Bibliography

Aristotle. *The politics*. Trans. J.E. Sinclair. Harmondsworth: Penguin Classics, 1962.

Baran, P.A. 1957. *The political economy of growth*. Harmondsworth: Penguin Books, 1973.

Bentham, J. 1793–5. *Manual of political economy*. In *Jeremy Bentham's economic writings*, ed. W. Stark. London: George Allen & Unwin, 1952.

Black, R.D.C. (ed.). 1977. *Papers and correspondence of William Stanley Jevons: Correspondence 1873–78*. London: Macmillan for the Royal Economic Society.

Cairnes, J.E. 1875. *The character and logical method of political economy*. London. Reprinted, New York: Kelly, 1965.

Cannan, E. 1929. *A review of economic theory*. London: P.S. King & Son.

Cantillon, R. 1755. In *Essay on the nature of commerce in general*, ed. H. Higgs. London: Macmillan, 1931.

de Mayerne-Turquet, L. 1611. *La Monarchie Aristodémocratique; ou le Gouvernement composé et meslé des trois formes de légitimes républiques*. Paris.

de Montchŕetien, A. 1615. In *Traité de l'économie politique*, ed. Th. Funck-Brentano. Paris: Plon, 1889.

Dobb, M.H. 1937. *Political economy and capitalism*. London: G. Routledge & Sons.

Engels, F. 1859. Karl Marx's 'A contribution to the critique of political economy'. *Das Volk*, Berlin, No. 14, August. In Marx (1859), 218–222.

Fraser, L.M. 1937. *Economic thought and language*. London: A. & C. Black.

Groenewegen, P.D. 1985. Professor Arndt on political economy: A comment. *Economic Record* 61: 744–751.

Hearn, W.E. 1863. *Plutology*. Melbourne: Robertson.

Hutcheson, F. 1755. *A system of moral philosophy*. Glasgow: Robert and Andrew Foulis.
Jevons, W.S. 1879. *The theory of political economy*, 2nd ed. London: Macmillan; Preface in 4th ed, London, 1910.
Jevons, W.S. 1905. *Principles of economics*. London: Macmillan.
Keynes, J.M. 1921. Introduction to Cambridge Economic Handbooks. In *Money*, ed. D.H. Robertson. London/Cambridge: Cambridge Economic Handbooks.
King, J.E. 1948. The origin of the term 'political economy'. *Journal of Modern History* 20: 230–231.
Knight, F.H. 1951. Economics. In *On the history and method of economics*, ed. F.H. Knight. Chicago: University of Chicago Press, 1963.
MacLeod, H.D. 1875. What is political economy? *Contemporary Review* 25: 871–893.
Marshall, A. 1890. In *Principles of economics*, 9th variorum ed, ed. C.W. Guillebaud. London: Macmillan, 1961.
Marshall, A., and M.P. Marshall. 1879. *The economics of industry*. London: Macmillan.
Marx, K. 1859. *A contribution to the critique of political economy*. Introduction by M. Dobb. London: Lawrence & Wishart, 1971.
McCulloch, J.R. 1825. *Principles of political economy with sketch of the rise and progress of the science*. London: Murray, 1870.
McKenzie, R.B., and G. Tullock. 1975. *The new world of economics: Explorations into the human experience*. Homewood: Irwin.
Mill, J. 1820. *Elements of political economy*, 3rd ed. London, 1926. Reprinted in James Mill, *Selected writings*, ed. D. Wisen. Edinburgh: Oliver & Boyd for the Scottish Economic Society, 1966.
Mill, J.S. 1831–3. On the definition of political economy; and on the method of investigation proper to it. Essay V in J.S. Mill, *Essays on some unsettled questions of political economy*, LSE reprint. London, 1948.
Mill, J.S. 1848. Principles of political economy with some of their applications to social philosophy. In *Collected works of John Stuart Mill*, ed. J.M. Robson. Toronto: University of Toronto Press, 1965.
Marquis de Mirabeau, V.R. 1758–60. *L'ami des hommes ou traité de la population*. Avignon/Paris.
Mishan, E.J. 1982. *Introduction to political economy*. London: Hutchinson.
Myint, H.L.A. 1948. *Theories of welfare economics*. London: Longmans, Green & Co.
Petty, Sir W. 1683. Observations upon the Dublin Bills of mortality and the state of that city. In *Economic writings of Sir William Petty*, ed. C.H. Hull. Reprinted, New York: Kelley, 1963.
Petty, Sir W. 1691. The political anatomy of Ireland. In *The economic writings of Sir William Petty*, ed. C.H. Hull. New York: Kelley, 1963.
Rees, A. 1968. Economics. In *International encyclopaedia of the social sciences*, vol. 4, ed. D.L. Sills, 472–485. New York: Macmillan.
Robbins, L. 1932. *An essay on the nature and significance of economic science*. London: Macmillan. 2nd ed, 1935.
Robbins, L.C. 1939. *The economic basis of class conflict and other essays in political economy*. London: Macmillan.
Robinson, J.V. 1933. *The economics of imperfect competition*. London: Macmillan.
Samuelson, P.A. 1955. *Economics*, 3rd ed. New York: McGraw-Hill. 7th ed, 1967.
Schumpeter, J.A. 1954. *History of economic analysis*. London: George Allen & Unwin.
Senior, N.W. 1836. *An outline of the science of political economy*. London: Unwin Library of Economics, 1938.
Smith, A. 1763. In *Lectures on jurisprudence*, ed. R.-L. Meek, D.D. Raphael, and P.G. Stein. Oxford: Oxford University Press, 1978.
Smith, A. 1776. In *An inquiry into the nature and causes of the wealth of nations*, ed. R.H. Campbell and A.S. Skinner. Oxford: Oxford University Press, 1976.
Steuart, J. 1767. In *An inquiry into the principles of political economy*, ed. A.S. Skinner. Edinburgh/London: Oliver & Boyd for the Scottish Economic Society, 1966.
Torrens, R. 1819. Mr Owen's plans for relieving the national distress. *Edinburgh Review* 32: Article XI.
Verri, P. 1763. Memorie storiche sulla economia pubblica dello stato di Milano. In *Scrittori Classici Italiani di Economia Politica*, Parte Moderna, vol. XVII. Milan, 1804.
Verri, P. 1771. *Reflections on political economy*. Trans. B. McGilvray, ed. P. Groenewegen. Reprints of *Economic classics*, Series 2, No. 4. Sydney: University of Sydney, 1986.
Young, A. 1770. *Rural oeconomy, or essays on the practical parts of husbandry*. London.

Political Economy and Psychology

P. H. Wicksteed

If political economy is the science of wealth, then it deals with efforts made by man to supply wants and satisfy desires. 'Want', 'effort', 'desire', 'satisfaction', are each and all psychic phenomena.

It would therefore appear that psychology must be to political economy – like the deity of Boethius – 'path, motive, guide, original, and end'.

Yet it is obvious that the political economist as such is not engaged in the establishment of the ultimate principles of psychology. He has not, for example, to investigate the nature of a concept, or determine the relation of the Will to the Reason. So far it is clearly true (cf. J.N. Keynes, *Scope and Method of Political Economy*, pp. 87, 88) that although the laws of the political economist 'rest ultimately upon a psychological basis', he accepts psychological principles as his data rather than establishes them as his conclusions; unless indeed he should be compelled to make excursions into the psychological field proper, because he does not find his premises sufficiently elaborated to his hand.

But this does not justify the reduction of the psychological factor of political economy to a level with the physical factor. Cairnes indeed (*Logical Method of Political Economy*, 2nd edn, pp. 37 and 38, quoted and apparently endorsed by Keynes, p. 85) instances the law of rent, and maintains that, in establishing this law, the economist no more undertakes to analyse the motives of self-interest which dictate the conduct of the landlord and the tenant than he undertakes to analyse the physical qualities of the soil which determine the law of decreasing returns. Now this is very true. The economist starts with both psychological and physical data, which he need not analyse, provided he has satisfied himself that they are true. But the difference is this, that whereas his *data* are partly physical and partly psychical, his *quaesita* are, in the last resort, wholly psychical. For if the law of rent is anything, it is a formulating of the principles which we may expect to regulate the *conduct* of men, secured in certain possessions and privileges, actuated by certain motives, and in the presence of certain physical facts and laws. The laws of political economy then, being ultimately laws of human conduct, are psychical and not physical; and therefore psychology enters into political economy on something more than equal terms with physical science and technology.

It therefore seems clear that, although the economist, as such, is not concerned with the ultimate analysis of his psychological data, his quaesita or conclusions are themselves of the order of psychic phenomena. But within the limits thus laid down there is still ample room for diversity of opinion. It may be contended that the economist has to receive, and test, his psychological and physical data alike, to deal with them by the universal methods of dialectic (i.e. inductive and deductive logic, or mathematics, if applicable), and then hand over his psychological results to the sociologist. Or it may be argued that political economy is largely, or even prevailingly, *applied psychology*, so that the economist must from first to last realize that he is dealing with psychological phenomena, and must be guided throughout by psychological considerations. In that case the relation of psychology to economics will be as close as that of mathematics to mechanics, though not in all respects analogous to it.

It is easy to see that the controversy as to the inclusion or exclusion of Consumption as a separate and acknowledged division of political economy, has a decisive bearing upon this question. The whole theoretic study of consumption can be little else than the application of the great psychological law of diminishing returns of satisfaction or relief to successive increments of commodity or service supplied to the same subject. To admit 'consumption' then as a branch of political economy is to admit that applied psychology has its conspicuous place in the science. So that if we are justified in saying that the express study of 'consumption' has now been definitively admitted as within the scope of political economy, we are thereby admitting psychological method, as well as psychological data and conclusions, as a part of the science; and the importance of dealing thus expressly with consumption and the psychological phenomena on which this branch of this study rests may well be shown by typical popular fallacies. For instance, there was no more common application of political economy a few decades back than the dictum that) 'what people want they will pay for', and that therefore all subsidizing is a waste of effort, and is 'against political economy'. Here the datum is that if *one and the same man* wants A as much as he wants B, he will be willing to give as much for it, sooner than go without it. From this datum certain conclusions as to market values and the commercially wise direction of efforts and resources are reduced, and these in their turn are reinterpreted into the statement that if *one* of two men is unwilling to give as much for A as *the other* is willing to give for B, then the first man does not want A as much as the second wants B, and it would be a wasteful and mistaken philanthropy to supply No. 1 with A rather than No. 2 with B. Of course no economist would formulate such an absurdity, but if the economists exclude consumption from express and psychological treatment, they leave room for and almost invite such 'applications'.

So much then for 'consumption'. But Exchange is so closely connected with consumption, and the laws of value are now seen to be so intimately dependent upon the psychological law of diminishing returns of satisfaction, that it must be impossible henceforth to exclude applied psychology from the problems of value and of exchange.

An excellent illustration is furnished by the problems of the currency. Of all branches of economic enquiry those that are concerned with Money and with Foreign exchanges seem most nearly to approach the objectivity of natural phenomena; and what is known as the Quantity Theory has been cited as a proof case of an economic law which is not psychological. But the truth is that no single step can be safely made in monetary science, unless the investigator keeps himself in conscious touch with his psychological basis during his whole investigation. We cannot, without special examination, even say that, in virtue of the universal law of supply and demand, the more sovereigns there are the lower will be their exchange value. For in this universal law of supply and demand there is a psychological link. Why does an increased supply lower exchange value? Because an increased supply of any commodity satisfies the corresponding want more completely, and reduces the unsatisfied remaining want to a lower degree of importunity. Now in the case of money it is admitted that within wide limits the money function is exactly as well performed by x and by nx pieces, so that there is no unperformed money function and money want becoming less and less importunate for satisfaction as the number of sovereigns, but not the command of commodities in general, increases. Thus, if the law of demand and supply is regarded as objective and absolute, and the psychological link forgotten, its applications to monetary problems will have no demonstrative cogency.

We now turn to Production and Distribution, and here we note at once that the study of 'production' must include the theory of labour, in which everything turns upon the law of the increasing irksomeness of successive increments of effort, and the decreasing psychological value of successive increments of commodity, or other result of effort; and the same law invades the study of distribution at every point, allying itself with the better-known physical law of diminishing returns to successive increments of any one factor of production, the others remaining constant.

In all the four main divisions of political economy, then, we see that the direction taken by economic study in recent years tends to a more express and generous recognition of the close connection between psychology and political economy, and the necessity of constantly keeping in touch with our psychological basis even when pursuing those branches of economic inquiry which appear to be remotest from it.

But, especially in connection with 'production' and) 'distribution', another aspect of the question forces itself on our attention. We have hitherto enquired whether the psychological data of economics can be accepted absolutely as results and dealt with by general dialectic methods, or whether they can only be considered as principles, to be applied with constant reference to the psychological conditions of the special problem under investigation. We have now to ask further, are these psychological data, whether facts or principles, to include all the psychological considerations that actually bear upon the production, distribution, etc. of wealth, or are we artificially to simplify our psychology and deal only with the motives supposed to actuate the hypothetical) 'economic man'? In the latter case political economy will be a hypothetical science. In the former it will aim at positivity.

And here again it will hardly be doubted that the tendency of recent work has been in the direction of enlarging the psychological area from which the data of political economy should be drawn. This tendency is manifested in two characteristic movements in recent economic investigation, which have in their turn reacted upon it. Firstly, the field of economic study, like so many others, has been invaded by the passion for the concrete method of enquiry, whether applied to contemporary or remote conditions. Now the man who studies the history of a great strike or trade-movement in Europe or America, of the land tenure or village industries of India, of middle-class or artisan budgets in England or France, of the growth and organization of industry in the Hanseatic cities or the republics of Italy, of the fiscal systems of commercially related peoples, and so forth, finds himself studying the conditions of the production and distribution of wealth, but in a region in which the simplified psychology of Ricardo and Senior is wholly inadequate. So conspicuously is this the case that some economists are ready to admit that no general theory or science of economics is possible, but only a natural history of wealth, production, etc., while others are seeking to reconstruct the general theory of economics on broader and more universally applicable principles. And it is here that the second movement characteristic of recent times allies itself with the historical method. It is the much-discussed mathematical method, which from this point of view is the necessary complement of the historical or concrete method. For no sooner has the mathematical student given to the acknowledged psychological data of economics the form, at once rigorous and generalized, that his method demands, than he perceives that his formulae really embrace the general theory of the distribution of resources with a view to maximizing a desired result, independently of the nature alike of the resources and the result in question. This brings the economic conduct of man under the same laws as his conduct in general, and promises to give us the wider basis of which we are in search.

Our conclusions throw a curious light on the much-debated but little-understood contention of Auguste Comte (*Philosophical Positivism*, vol. iv. pp. 193 et seq.) that there is no specific science of wealth, with special laws and principles, and that the attempt to deal with the wealthgetting impulses of man in isolation must be essentially barren; but the special applications of general principles of philosophy to the industrial and commercial life may be prolific and illuminating in a high degree.

Bibliography

Cairnes, J.E. 1875. *The character and logical method of political economy*, 2nd ed. London: Macmillan.

Comte, A. 1830–42. *Cours de philosophie positive*. 6 vols, Paris: Bachelier.

Keynes, J.N. 1891. *The scope and method of political economy*. London: Macmillan.

Political Economy Legacy of Institutions from the Classical Period of Islam

Lisa Blaydes and Eric Chaney

Abstract

This article describes the core political and economic institutions of Muslim societies during Islam's 'classical' period. We argue that the reliance of Muslim leaders on slave armies discouraged the development of a hereditary baronage in Muslim societies and contributed to the underdevelopment of private ownership of land. Societal resistance to sultanistic governance emanated not from land-owning elites but rather from religious leaders who came to enjoy high levels of moral authority through their role as mediators between state and society. Authoritarian governance, a weak tradition of private property rights and empowerment of religious elites in the realms of law and education had important implications for the economic development of Muslim societies.

Keywords

Economic development; Institutions; Political economy; Political power

JEL Classifications

N45; O10

Why is the Islamic world underdeveloped relative to other world regions? This question is particularly puzzling when one considers the region's historical leadership in fields as diverse as commerce and science. Over the past decade, a body of research has highlighted how Islamic legal institutions contributed to the region's underdevelopment (Kuran 2001, 2004, 2011). We argue that while Islamic legal institutions can be usefully viewed as one proximate cause of underdevelopment, the Islamic world's unique political equilibrium is the fundamental cause of its poor economic performance over the past centuries. We also explain how and why this political equilibrium materialised in addition to highlighting its impact on the political and economic development of Muslim societies.

The Emergence of the Classical Institutional Equilibrium

Following the death of the Prophet Muhammad in 632 CE, Arab armies conquered large swathes of West Asia and North Africa. Initially, the Arab-Muslim conquerors maintained the well-developed administrative bureaucracies of their conquered predecessors and largely lived in garrisons separate from the local populations. Eventually, Muslim rulers began to introduce new institutional forms for the maintenance of political order. The most important of these innovations was the introduction of slave armies by the Abbasid Caliphs in the ninth century and the diffusion of such armies across the dynasties of the medieval Islamic world. Historians have viewed the introduction of these armies as a 'major innovation in Middle Eastern history' (Lapidus 2014, p. 86) and their appearance is believed to have contributed to political instability in the region (Blaydes and Chaney 2013).

The widespread use of slave armies transformed Muslim societies in at least two ways. First, ruler reliance on slave soldiers who had been imported from non-Muslim lands meant that local holders of military power lost influence and were eventually eliminated as a social force. Second, as local military elites saw their power undermined, religious leaders emerged as the primary representative of societal interests and a key 'check' on executive power. Over time, with the growing conversion of local populations to Islam, Muslim religious leaders began to take 'charge of Middle Eastern communities' and to 'infuse them with their interpretation of Islamic identifications' (Lapidus 2014, p. 224). We call the political equilibrium which emerged as a result of these transformations the 'Classical Institutional Equilibrium' (henceforth, CIE). At its core, the CIE reflected a division of political power – broadly speaking – between Islamic religious leaders and Muslim kings reliant on slave armies, or in some cases slave armies without kings, as in Mamluk Egypt.

While elements of the CIE eventually spread to much of today's Islamic world, in our view the CIE best characterises regions incorporated into the Islamic world during the medieval period. In addition, we believe that although this political equilibrium emerged at different times in different regions, generally speaking it was fully developed by the twelfth century.

The Classical Equilibrium in Action

Thus far, we have argued that the institutions associated with the classical equilibrium reflect the interests of Muslim rulers dependent on foreign-born slave soldiers and Muslim religious elites. The social and political arrangements between these two groups reflected their collaboration on the one hand and the tension between them on the other. In this section, we describe the institutions generated by the CIE in two domains: property rights and human capital formation. In both cases, we highlight how outcomes reflect an accommodation between the two social groups at the core of the CIE.

Property Rights, Predation and the Waqf

Secure private property rights are seen as key to encouraging growth-producing investment. The system of property rights which predominated in the CIE may provide a partial explanation for why Muslim societies lagged behind Europe and other world regions in the development of those rights.

The widespread use of foreign-born slave soldiers was closely linked to forms of fiscal organisation which discouraged the emergence of secure property rights. In order to pay for these elite soldiers and to reduce their incentive to rebel, Muslim rulers allotted control over government-owned land or other rent streams to slave soldiers as compensation. In exchange for their loyalty and maintenance of social order in that territory, slave soldiers had the right to tax revenue from that land. In contrast to developments in medieval Europe, such assignments were generally short-term and not intergenerationally transferable. They were also revocable and exchangeable by order of the government (von Grunebaum 1970, p. 145). As a result, these land grants were qualitatively different from the feudal fiefs of medieval Europe, as slave soldiers rarely established enduring roots to the land and were often rotated from locale to locale. Given the relatively short time horizon they envisioned in any particular location, it is not surprising that land grant holders were more concerned with extraction of rents and less concerned with establishing secure property rights and other long-term investments in the areas to which they were assigned. (Olson (1993) would suggest, however, that Muslim sultans and their slave armies retained a sufficient interest in maintaining the productivity of land assets as to avoid the uncoordinated and competitive theft associated with 'roving' bandits.)

In the context of predation by rulers and their associated slave armies, there emerged a societal need for an institution to shelter wealth for those able to accumulate capital. It is in this context that we believe that the spread of *waqf*, or Islamic charitable trust, must be understood. Scholars have long argued that wealthy individuals used the *waqf* to both protect assets from predation and to better control how their assets would be divided following their death and thus escape the reach of restrictive Islamic inheritance laws.

*Waqf*s were under the jurisdiction of Islamic religious leaders who emerged as a key holder of domestic political power and a representative of societal interests in the CIE. In order to better serve their 'constituencies', religious leaders channelled many of the resources put into *waqf* towards public goods that the heavily militarised government in the CIE did not provide. These Islamic trusts often took the form of schools, rest houses for pilgrims, public baths, water fountains and hospitals. The *waqf* founder also benefitted from this arrangement. The founder was able to designate both the beneficiaries of the *waqf* as well as the compensated administrator (often himself or a member of his family). Why didn't Muslim rulers and their slave soldiers confiscate these Islamic trusts? If they had, religious leaders would have encouraged revolt against these regimes. Knowing this, the *waqf* remained relatively safe from predation.

Kuran (2001, 2004, 2011) has highlighted the importance of *waqf*s in his studies of economic development in the Muslim world. Kuran (2001, 2004, 2011) argues that the *waqf* immobilised assets in static perpetuity, creating significant inefficiencies. Because the founder of the *waqf* designated a particular purpose for the pious foundation at the time of its creation, rigidities built into this mandate became particularly dysfunctional over time.

While we agree with Kuran that the *waqf* tied up capital in ways that were bad for economic growth, we see the *waqf*, as an institution, as the outcome of the balance of political power in the CIE rather than an independent cause of economic stagnation. While the earliest *waqfs* likely date to late ninth and early tenth centuries, it was not until the twelfth century that the practice of 'of putting landholdings into the form of *waqf*, pious endowments, inalienable and not subject to government seizure, became common' (Hodgson 1974, p. 51). As it result, it is difficult to argue that the *waqf* was a direct, time-invariant derivative of Islamic doctrine. Instead, we see the growing usage of the *waqf* as the result of a political bargain struck between religious leaders, who emerged as a politically influential group as a result of the introduction of slave armies and rulers backed by these armies. This argument is consistent with scholars who focus on the political underpinnings of Islamic legal institutions (Malik 2012).

Religious Elites, Education and Human Capital Formation

Politically influential religious leaders could, to some degree, limit predation by rulers and their associated military elite; their influence did not end there, however. The emergence of the CIE coincided with the spread of educational institutions controlled by religious leaders, of which the *madrasa* is the best known. *Madrasa* education placed a premium on religious knowledge over other forms of learning. Over time, religious education crowded out the study of scientific subjects. In addition, the rise of this educational system is believed to have discouraged innovation and the development of broad forms of human capital.

Chaney (2015a) provides empirical evidence consistent with this hypothesis. In particular, he documents a decrease in scientific production that corresponds with the rise of *madrasa*s across the Islamic world. Such evidence complements work by historians noting that 'the institutionalization of Muslim scholarship' in *madrasa*s marked 'a significant change in Islamic social structure and Muslim community life' (Chamberlain 1994, p. 69).

The available evidence stresses that religious leaders sought to limit the study of some topics, particularly as related to science and mathematics, because they believed that certain forms of knowledge weakened their societal influence. For example al-Ghazali (1058–1111 CE), a well-known religious scholar, noted that 'he who studies mathematics is amazed by its precision and proofs. He then takes a more positive view of philosophy and reckons that all of the rational sciences are as clear and trustworthy as mathematics [. . .] and he says: if religion were true, then philosophers would have proved its veracity with their precise methods [. . .] we have seen many deviate from [Islam] in this manner' (al-Ghazali 1111 [1971], pp. 21–2). This suggests that scientific/innovative human capital formation decreased both the quality (through obedience) and quantity (through size of following) of a given leader's popular support. If true, this implies that religious leaders should work to limit the types of human capital formation that undermined their population-based forms of political power (Acemoglu and Robinson 2000; Chaney 2013). The net result was that after the emergence of the CIE research became increasingly limited to 'very narrow, and essentially unprogressive areas', as religious leaders, instrumentally, sought to limit the permitted areas of scholarly focus (Sabra 1987, pp. 240–1).

The decline in scientific output in the Muslim world contrasts sharply with the well-known increased production of scientific knowledge in Western Europe during the late medieval period (e.g. the Renaissance). This increase in European scientific production coincides with a decline in the focus on religious discourse in European political theory (Blaydes et al. 2016). Taken together, such evidence suggests that relatively high rates of human capital formation in Europe may be due – at least in part – to the greater constraints faced by European religious elites in the later medieval period.

The Long-Run Impact of the CIE on Political Institutions

Have the political institutions of the CIE had an enduring impact? We argue that insecure property rights and low levels of human capital formation damaged the economic and technological development in CIE regions, paving the way for European colonialism. There is little doubt that by the end of the seventeenth century Muslim societies increasingly struggled to keep pace with technological developments in Western Europe (Chaney 2015b). This represents a meaningful reversal when compared to the medieval period, when Western Europe clearly lagged behind the Islamic world on these dimensions.

Beyond growth-related outcomes, the CIE also impacted political institutions. Chaney (2012) provides evidence that the Islamic world's 'democratic deficit' is driven by areas which were subject to the Arab conquests and thus typified the CIE. The key contrast within the Muslim world, then, is between those countries – like Iran, Egypt and Syria – which Arab armies conquered in the early expansion of Islam versus Muslim-majority countries beyond the Arab conquest regions – like Bangladesh, Indonesia and Senegal – which exemplified the CIE to a lesser extent. Based on this evidence, Chaney (2012) suggests that the Islamic world's democratic deficit has less to do with Islam *per se* than the long-run impact of the CIE.

The political role of religious leaders in CIE regions today appears to have parallels with the historical experience. In particular, religious leaders, organisations and political parties are dominant contemporary social actors in Muslim societies. These organisations seem to exert high levels of political power, at least when compared to their counterparts in other world regions. This relatively high level of political influence has a significant impact on institutional development.

For example, Blaydes and Lo (2012) argue that democratisation in the Middle East has been hampered by the fact that religious associations are typically the best-organised and most important civil society groups. Because potential regime liberalisers in Middle Eastern autocracies suspect political openings could become a vehicle for Islamists to seize power through free elections, Middle Eastern autocrats are reluctant to provide openings for societal organisation. (This argument contrasts with scholars who suggest that Middle Eastern autocrats strategically cultivate Islamist opponents in a bid to force citizens into supporting status quo dictatorship versus religious extremists.) This suggests that the dominant position of religiously motivated civil society has decreased the possibility of political liberalisation in the Middle East. On this dimension, Middle Eastern societies differ from newly democratising states in Eastern Europe, Latin America and East Asia where civil society elites have typically been secular liberals.

See Also

- ► Economic Growth
- ► Human Capital
- ► Islamic Economic Institutions
- ► ‘Political Economy’
- ► Political Institutions, Economic Approaches to
- ► Power
- ► Religion and Economic Development
- ► Political Economy of Institutional Change: Illustrations from the Ottoman Empire

Bibliography

Acemoglu, D., and J.A. Robinson. 2000. Political losers as a barrier to economic development. *American Economic Review* 90(2): 126–130.

Al-Ghazali, A.H. 1111 [1971]. *al-Munqidh min al-Dalal*. Lahore: Hayat al-Awqaf.

Blaydes, L., and E. Chaney. 2013. The feudal revolution and Europe’s rise: Political divergence of the Christian West and the Muslim world before 1500 CE. *American Political Science Review* 107(1): 16–34.

Blaydes, L., and J. Lo. 2012. One man, one vote, one time? A model of democratization in the Middle East. *Journal of Theoretical Politics* 24(1): 110–146.

Blaydes, L., J. Grimmer, and A. McQueen. 2016. *Mirrors for Princes and Sultans: Advice on the art of governance in the Medieval Christian and Islamic worlds*. Working paper.

Chamberlain, M. 1994. *Knowledge and social practice in Medieval Damascus*. Cambridge: Cambridge University Press.

Chaney, E. 2012. Democratic change in the Arab world, past and present. *Brookings Papers on Economic Activity* 42(1): 363–414.

Chaney, E. 2013. Revolt on the Nile: Economics shocks, religion and political power. *Econometrica* 81(5): 2033–2053.

Chaney, E. 2015a. *Religion and the rise and fall of Islamic Science*. Working paper.

Chaney, E. 2015b. Measuring the military decline of the Western Islamic world: evidence from Barbary ransoms. *Explorations in Economic History* 58: 107–124.

Hodgson, M. 1974. *The venture of Islam*. Chicago: University of Chicago Press.

Kuran, T. 2001. The provision of public goods under Islamic law: Origins, impact, and limitations of the waqf system. *Law & Society Review* 35(4): 841–898.

Kuran, T. 2004. Why the Middle East is economically underdeveloped: Historical mechanisms of institutional stagnation. *Journal of Economic Perspectives* 18(3): 71–90.

Kuran, T. 2011. *The long divergence: How Islamic law held back the Middle East*. Princeton: Princeton University Press.

Lapidus, I.M. 2014. *A history of Islamic societies*. 3rd ed. New York: Cambridge University Press.

Malik, A. 2012. *Was the Middle East’s economic descent a legal or political failure? Debating the Islamic law matters thesis*. Working paper, Center for the Study of African Economies.

Olson, M. 1993. Dictatorship, democracy, and development. *American Political Science Review* 87(3): 567–576.

Sabra, A.I. 1987. The appropriation and subsequent naturalization of Greek science in medieval Islam: a preliminary statement. *History of Science* 25: 223–243.

von Grunebaum, G.E. 2008 [1970]. *Classical Islam: A history, 600 AD to 1258 AD*. New Brunswick: Transaction Publishers.

Political Economy of Arab Uprisings

Adeel Malik

Abstract

This article frames the political economy of the 2011 Arab uprisings as a failure of the Arab development model, especially its inability to support an independent and competitive private sector. Based on a distorted legacy of intervention and distribution, this development model is structurally incapable of reconciling aspirations with economic opportunities. The contradictions associated with this development model are particularly apparent in the region's labour-abundant economies, where a shrinking resource envelope has led to an erosion of the social contract, resulting in a scaling back of public employment and welfare services. Worryingly, the space vacated by a shrinking state has not been filled by a vibrant private sector. This article analyses the crisis of the Arab state through the lens of an under-developed private sector. In much of the Arab world the private sector acts as an appendage of the state. Businesses tend to survive either when they are too close to the state, such as crony capitalists, or too far, which is the case with informal firms. While private sector development remains an important imperative, it is not simply a function of technocratic policy reform. Relieving greater competitive space for the private sector requires a political concession that grants autonomy to independent businesses and relaxes barriers to regional trade. I argue that an independent merchant class is difficult to visualise without connected regional markets.

Keywords

Arab economies; Private sector; Regional economic cooperation; Rents

JEL Classifications

O10; O53; P26; R11

The Arab revolutions that started in December 2010 in Tunisia and quickly spread across the region had a clear economic underpinning. They were fuelled by poverty, unemployment and lack of economic opportunity. At their heart was a yearning for social justice. While political repression in the Middle East remains a subject of regular debate, the scale and intensity of the region's economic repression has gone relatively unnoticed. The Middle East has long been trapped in a vicious development cycle, defined by an excessive dependence on natural resources, a Leviathan state and a weak private sector. This has prevented the rise of a strong constituency for economic diversification. This article argues that continuing political upheaval in the Middle East cannot be understood without fathoming the growing unsustainability of the Arab development model. Underscoring an inherent tension between the region's demographic and economic structures, I argue that, while the Middle East has undergone an unprecedented demographic transition, its economic structure remains rigid, unable to generate productive employment opportunities for new entrants to the labour force. The profound employment challenge that Arab societies face today cannot be met without conceding more competitive space to the private sector. But this is determined by political, not just economic, choices. Arab regimes face a difficult politics of private sector development defined by an inherent trade-off between employment and autonomy. While rulers wish to promote employment generation, they are unwilling to cede greater autonomy to the private sector.

An Evolving Demography, but a Rigid Economic Structure

A growing tension has developed in the Middle East between demography and economic structure. The Arab Spring has brought into sharp focus the profound implications of a demographic shift, whereby an overwhelming proportion of the region's population – in many countries about 75% – consists of young people under the age of 30. A significant proportion of this young population is also female and educated. While Arab societies have failed on several development fronts, they have had a resounding success in expanding access to education. Challenges of educational quality aside, many Arab countries, especially those in North Africa, have made impressive strides in educating their young and closing the gender gaps in education. In fact, of the top ten countries that have made tremendous progress in human development during the last 40 years, five were from the Arab world. The key problem, however, is that education is not matched with economic opportunity. While the public sector could absorb all new entrants to the labour force 40 years ago, it is stretched beyond its limits in many labour-abundant Arab countries today, resulting in some of the world's highest rates of youth unemployment. While education has enhanced aspirations unemployment has only resulted in grievance.

Many of these young people are not only unemployed, they are also *unemployable*. This is clearly a failing of both the education system and economic structure. Educational institutions churn out graduates who have credentials that make them eligible for public sector jobs, but skills that are largely irrelevant for the private sector. Given better wage remuneration and job security in government employment, the young are dissuaded from pursuing a career in the private sector. In many countries (especially in the oil-rich Gulf) this leads to a perverse division of labour between the public and private sectors. Typically, the public sector generates high-wage jobs for nationals, while the private sector is overwhelmingly reliant on expatriate workers. Such labour market segmentation has profound implications for political economy, since it prevents both citizens and the state from developing a genuine stake in private sector development.

The region's profound demographic shifts need not be a liability. As the East Asian experience suggests, they can be harnessed for growth and serve as a demographic dividend. The irony in the Middle East, however, is that, while demography is evolving, the economic structure remains unresponsive to the needs of its growing populations. With the public sector still acting as the employer of first resort, the Arab world suffers from a precarious employment strategy. In many of the region's labour-abundant economies this is further complicated by a weakening of the social pact that had initially ensured employment for all, but which is now growingly incapable of absorbing new entrants to the labour force (Cammett and Diwan 2014). This retreat of the welfare state has not been accompanied by the development of a strong private sector capable of picking up the slack in job creation. The private sector – despite its varying strength across Arab economies – remains weak and dependent on state patronage. Its failure ultimately emanates from a development model whose expiry date has long passed, but is being sustained through the regular injection of externally derived rent streams.

A Failed Development Model

In most Arab economies the state has typically remained the most important economic actor, eclipsing all other productive sectors. When it comes to the essentials of life – whether food, jobs, housing or public services – the state is the provider of first and last resort. The functioning of this system rests on a heavy dose of subsidies and economic controls, and a variety of other uncompetitive practices. While a centralised bureaucratic system has worked well for ruling elites and the narrow clienteles that thrive with their support, it has failed to deliver prosperity and social justice to ordinary citizens and firms. The interests of governing coalitions have proved more enduring than the force of ideology. Neither the socialism of the 1960s and 1970s nor the neo-liberal economic reform of the 1990s has been able to dismantle this system of centralised control, discretion and privilege. This development model is structurally incapable of reconciling aspirations with opportunities.

This state-centred development paradigm rests on the uninterrupted flow of external windfalls. In fact, many of the region's pathologies – whether a weak private sector, segmented labour markets or limited regional trade – are ultimately rooted in an economic structure that relies heavily on external rents, whether derived from fuel exports, foreign aid or remittances. Reliance on these unearned income streams is the ‘original sin’ of Arab economies. More than 80% of total merchandise exports in many Arab countries consists of oil and gas. The dependence on hydrocarbons is so pervasive that even in economies otherwise considered to be resource-scarce, such as Syria and Yemen, exports are dominated by oil. Up until 2005, for example, around 67% of the total exports in Syria consisted of fuels. In Yemen, fuel exports constitute 70% of total exports.

Where oil is relatively scarce, foreign aid replaces its role. Aid revenues, much like oil, tend to stifle economic and political incentives, turning economies away from production to patronage. By virtue of their strategic locations, Egypt and Jordan derive significant external rent streams through foreign aid. In Egypt alone – hardly a typical case of resource curse – two-thirds of foreign exchange revenues are derived from oil, aid and revenues from the Suez Canal. While the role of oil dominates the discourse on the Middle East, the influence of aid is often downplayed. It may come as a surprise that, as a region, the Middle East and North Africa (MENA) received the highest overseas development assistance on a per capita basis in 2008 ($73 compared to $49 in sub-Saharan Africa). Resource windfalls from oil and aid are complemented with remittances. As a ratio of GDP, the MENA region received the highest proportion of remittances. The external revenues from oil, aid and remittances sustain an adverse political economy predicated on a social pact that trades welfare distribution for regime security.

While these unearned income streams continue to finance the Arab social contract, the resource envelope in many Arab economies is not expanding as fast as the cost of the underlying social bargain. This is particularly apparent in economies that are labour-abundant but scarcely endowed with natural resources. While, initially, oil revenues in these economies were fiscally important, they have declined precipitously since the 1980s. This has been accompanied by a withdrawal of the state’s welfare functions and followed by a malign process of economic reform that replaced public monopolies with private cronies. The numbers are indicative of this dramatic shift. Since the 1980s oil revenues have more than halved in Egypt and Tunisia, paving the way for a significant downsizing of the state. The ratio of public expenditures to GDP in these countries has declined from 50% in 1980s to 30% by the early 1990s. The decline was particularly dramatic in Egypt, where the ratio fell from 61.5% in 1982 to 25.1% of GDP in 1998 (Diwan 2014).

This fiscal retreat of the state was accompanied by a decline in the quantity and quality of public spending. Public investment has borne the major brunt of expenditure cuts. Resources allocated for salaries and subsidies have declined. Notwithstanding the recent hike in subsidies, expenditures earmarked for subsidies originally fell from 9.7% in 1980s to around 1.1% of GDP in the 1990s (Diwan 2014). Such fiscal contractions have weakened the Arab social bargain in relatively resource-poor economies, leading to a scaling back of job opportunities and welfare services in health and education. While co-opted through initially generous welfare entitlements, the Arab middle classes, composed largely of state civil servants, are now growingly disaffected with this gradual erosion of the social contract (Diwan 2013). But, as recent survey evidence has suggested, such middle class ‘grievance’ is not uniform across countries and generations (Cammett and Silti 2014). In some countries, for example, the young cohort is more dissatisfied with welfare services and labour market opportunities than the older generation.

Importantly, the grievance is not just about incomes: it is also about opportunities. As public sector jobs decline, they are rationed by connection rather than competition. The welfare regime therefore faces the twin challenge of access and fairness. With the emasculation of the Arab welfare state, the burden has disproportionately fallen on the poor and the middle classes. The socioeconomic cost of this fiscal adjustment has heightened due to a weak institutional base. Rather than visualising a shared political settlement to compensate losers and achieve a more equitable sharing of the burden of adjustment, the ruling coalition has been further narrowed by restricting economic opportunity to the closed circle of family and friends. This was driven by a sound political logic. Authoritarian regimes need to placate both citizens and elites. Such appeasement is easier in oil-abundant economies, where pockets are deep enough to satisfy the two constituencies. In resource-poor economies, however, rents from government manipulation of the economy play a more crucial role in sustaining elite coalitions.

Connected elites in these countries are given control of vital access points to the economy. As long as these insiders monopolise the economic rent streams, their interests are aligned with the continuity of the regime. In many of the Middle East's labour-abundant economies state-led industrialisation offered precisely such opportunity to secure rents for the connected. But, as oil revenues fell and fiscal adjustment became a necessity, economic liberalisation shifted the rentier equilibrium. Liberalisation generated new rents for insiders through trade policy concessions, stakes in privatisation and new brokerage opportunities through partnership with foreign firms. Rather than a genuine levelling of the economic playing field, liberalisation simply gave rise to crony capitalism.

The emergence of such connected capitalism was most visible in Tunisia and Egypt. In Tunisia, Ben Ali and family ended up controlling 220 firms. Many of these firms operated behind high tariff and non-tariff barriers and were primarily active in sectors that required prior government authorisation and were subject to FDI restrictions (Rijkers et al. 2014). Recent evidence from Egypt suggests that politically connected firms operated behind high non-tariff barriers and had preferential access to subsidies and credit. In Egypt 71% of the connected firms were protected by at least three non-tariff measures relative to 3% of all firms included in the sample. Moreover, politically connected firms were four and a half times more likely to operate in energy-intensive sectors (Diwan et al. 2013). Even as the state contracted, releasing space for the private sector, it did not widen economic access to ordinary firms. Entry barriers remained intact, and large parts of the economy were still governed by licence requirements. Productive resources, such as land, were accessible only to a few.

Such insider advantage has profound social implications. While connected firms in Egypt accounted for 60% of net profits and 92% of loans, they accounted for only 11% of employment (Diwan et al. 2013). The unemployed usually ended up in unproductive subsistence activities in the informal sector. All across MENA's labour-abundant economies the bulk of job creation takes place in the informal sector. This generates an important political economy dilemma. While privileges are concentrated among politically connected firms, employment is concentrated among small informal sector firms. Since the onset of liberalisation, the default welfare model in the labour-abundant economies of MENA has been based on welfare provision for citizens and privileges for connected firms. While the system affords subsistence, the social mobility of both people and firms is compromised. This intensifies the divide between insiders and outsiders, traps educated populations into unproductive jobs and keeps social classes dependent and immobilised (Malik 2014).

With the shrinking resource envelope, it is no longer possible to accommodate the growing pool of educated unemployed in the public sector. As a result, the unconnected, young and females have become the new outsiders to labour markets (Assad 2014). Enterprises operating on the margins of the economy – whether small, medium or informal – suffer from institutionalised discrimination. For a long time the economics of concessions has trumped the economics of competition in the Middle East. To varying degrees, most Arab countries are witnessing a generational struggle for inclusion, where young people and firms are aspiring for mobility. Even in the resource-rich Gulf, where underlying social tensions are contained through regular resource injections, the ruling bargain that 'trades welfare distribution for political acquiescence' remains vulnerable (Malik and Awadallah 2013).

There is a growing realisation among ruling circles that the prevailing social contract is unsustainable in the face of rising energy demand at home, growing oil production from alternative sources, and fast expanding populations. For example: even with robust oil prices, Saudi Arabia is faced with long queues for public sector jobs. With 300,000 jobs needed annually to accommodate young Saudi graduates, the waiting period for public employment has increased tremendously (*Saudi Gazette* 2014). Changing demography, access to technology, growing food prices and a weakening distributive pact have scaled up the cost of repression and redistribution – the twin pillars of authoritarian order in the Middle East. Even as public expenditures have fallen in resource-poor economies, spending on repression (especially internal security) has remained both robust and resilient. For decades, the Arab state, regardless of whether it is a monarchy or a republic, has ruled through the fear of its security services. It has perfected the art of demolishing any commons imaginable. Social media have generated new spaces for collective action, however. These are the virtual commons that cleverly evade the long arm of the state.

In short, when faced with political shocks Arab regimes have typically fallen back on extended patronage commitments rather than strong institutions. Rulers in the resource-rich Gulf have tried to placate their populations with public sector jobs, salary increases, concessional loans, infrastructure contracts and one-off gifts. This can only purchase temporary stability. The Arab revolts of 2011 symbolised the yearning for a new social contract that could provide ladders for economic and social mobility – ladders that have been denied by the prevailing Arab development model. The region needs a new economic paradigm that is based on a competitive, entrepreneurial and dynamic private sector. But fostering such a private sector requires a new pattern of politics.

The Politics of Private Sector Development

The Arab Spring was not just a crisis of the Arab state – its inability to redistribute, reform and represent the interests of ordinary citizens. It was also a crisis of the private sector. The popular movements that overthrew Mubarak and Ben Ali were as much targeted against dictators as the crony capitalists surrounding the royal circle. In fact, the changing character of business–state relations provides an important window into understanding Arab political economy. As the state in labour-abundant Arab societies shrank, the space it vacated was not filled by the private sector. Rather, declining government employment was compensated by the informal sector. Recent micro evidence suggests that, as the share of workers employed by the government declined, the slack was largely picked up by the informal sector, which employed less than 5% of total Egyptian workers in 1970, but accounted for 40% of workers in 2005 (Assad 2014). By contrast, the share of employment in the private sector actually declined.

While private sector development is frequently recognised by official circles as essential for employment generation, both donors and national governments have viewed the underlying challenge largely through a narrow technocratic lens that is shorn of both history and politics. The problem is often framed as a result of defective policies, manifested through a high cost of doing business, a poor investment climate and weak indicators of global competitiveness. Developing a strong private sector is not simply about removing regulatory constraints, however. It is also a political and regional challenge. To the extent that an independent private sector can shape new patterns of economic and political power, it is considered a political threat by rulers. And there is also a strong regional dimension to private sector development. A crucial barrier to private sector growth is the persistent economic fragmentation of the Arab world into isolated geographic units that are, at best, weakly connected with each other. Connected Arab markets are crucial for affording scale economies to firms. These two dimensions are now considered separately.

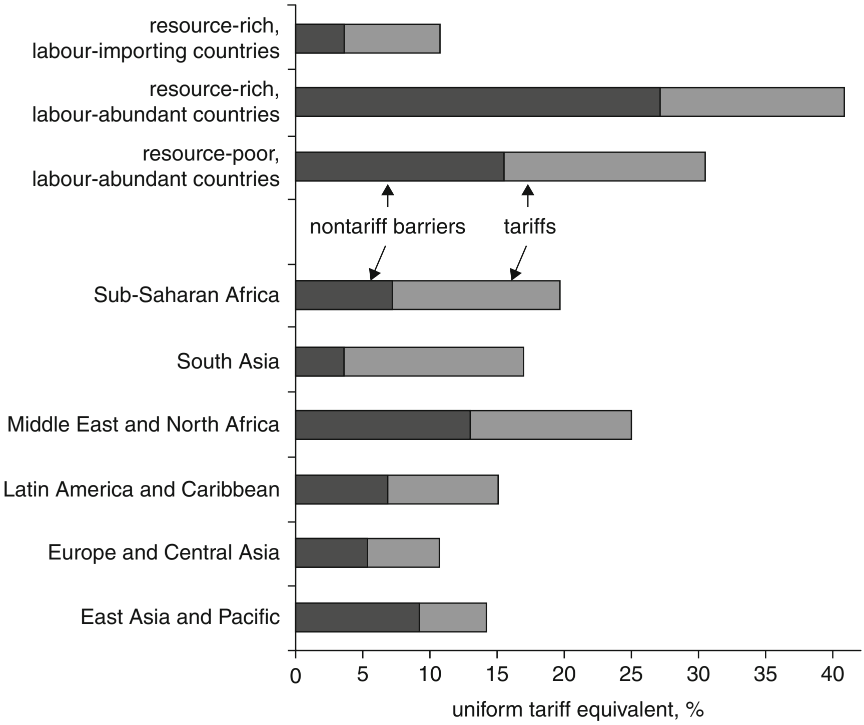
Amongst the list of constraints to private sector development, the role of politics stands out. There is growing evidence that the private economic domain in the Middle East is defined by privileges rather than competition. There is a sharp disjunction between the *de jure* and the *de facto*: even when productive activity is governed by laws, they are inconsistently applied, which indicates the inherently arbitrary and discretionary nature of economic policymaking. The private sector is usually a mirror image of the state: inefficient, controlled by a tiny clique of elite families tied to ruling regimes and part of an extensive network of patronage. Its profits depend less on entrepreneurial abilities and more on access to power. Exploiting new economic opportunities therefore becomes a game of insiders. With few exceptions, major business fortunes are accumulated through 'closed' deals with regime insiders.

Even in the Gulf, where the private sector is admittedly more vibrant, the boundaries between the public and the private are noticeably blurred. Rulers and businessmen are often indistinguishable. With the discovery of oil in 1950s the ruler–merchant relationship shifted firmly in the favour of rulers. Businesses are allowed to grow only so long as they remain loyal and dependent. Its impressive growth notwithstanding, the private sector lacks an autonomous political voice in Gulf countries (Hertog et al. 2013). Historical legacy offers some explanation as to why private enterprise has remained persistently weak in the Middle East. Independent sources of economic power have typically drawn fear rather than favour from Arab rulers. Merchants were largely absent from the power configuration in Ottoman Empire. Ottoman rulers were more inclined to grant economic concessions to European and minority merchants who were less likely to pose any direct threat to the Sultan's power. In many successor Arab states, independence was followed by the exodus of foreign merchants, leaving behind a vacuum that deprived the region of an important constituency that could have pushed for genuine economic reform. To make matters worse, whatever weak private economic activity survived after independence was nationalised by successor Arab states under the garb of socialism. This subjected merchants to relatively adverse initial conditions to begin with.

The Licence Raj established after independence has survived longer in the Middle East than elsewhere in the developing world. Today, severe restrictions remain in place on the movement of goods and labour across Arab borders. Although tariff barriers have been slashed, the more invisible and non-transparent behind-the-border barriers continue to be a source of trade frictions. The trade barriers that create economic enclaves for insiders are difficult to dismantle, since rents generated by such restrictions are used to manage ruling coalitions. In this sense, such barriers are not just procedural but also political barriers. This is an important reason why non-tariff barriers have remained more pervasive in MENA's labour abundant economies (see Fig. 1). It is consistent with the idea that, given their more extensive distributional commitments, labour-abundant economies tend to rely on alternative rents from government manipulation of the economy. Why would political incumbents, in this milieu, vacate space for new entrants when their survival depends on closing off economic access? If the incentive structure that governs private enterprise in MENA is the outcome of political choices, a simple insistence on technocratic reforms is unlikely to work. A competitive space for the private sector will therefore have to be negotiated as a political concession.

We next turn to the regional dimension. A core structural reason for the region's historically weak private sector lies in fragmented markets. Despite its rich trading past and common cultural heritage, the Arab world is one of the most divided regions in terms of productive economic linkages. In a region of 350 million people, mutual trade between Arab economies remains minimal, hovering around a paltry 10% of total merchandise trade. This is a colossal economic failure and has wider repercussions for the region than are commonly understood. The absence of a large connected market denies Arab firms economies of scale, which fuels both growth and diversification. The inability to produce for a larger market caps the growth potential of firms, reinforces the dependence of firms on state patronage and denies the private sector an opportunity to become an independent driver of socio-economic change.

But it is not simply about the absence of scale economies. Economic fragmentation prevents the emergence of regional supply chains – a key driving force behind global trade expansion. It also makes deeper trade reforms more



Political Economy of Arab Uprisings, Fig. 1 The politics of trade barriers in the Middle East (*source*: Shui and Walkenhorst 2010)

unlikely, since regional trade cooperation is often more helpful in identifying and dismantling non-tariff barriers. Thin markets also reinforce the political economy of protection, preserve the monopoly power of insiders and increase the returns to predatory behaviour. With divided markets, the market for second-hand capital goods remains underdeveloped, making new investments particularly risky as businessmen face the risk of being stuck with bad investments. Another cost – largely ignored in most literature – is the wasteful duplication of defence expenditures. As a region, MENA is the biggest global spender on defence (as a share of GDP). During the past decade, the region spent twice as much on defence as South Asia. Another cost of fragmentation is manifested in the under-provision of regional public goods. Connective regional infrastructure is one such public good. Since the benefits of regional transport infrastructure are likely to accrue to everyone in the region, individual countries are discouraged from investing in it, which generates a massive coordination failure.

The Arab world's economic fragmentation is puzzling, given its favourable geography. The MENA region lies at the inter-section of major trading routes, with easy access to markets in Africa, Asia and Europe. And, while, in Africa nearly 40% of the population lives in landlocked countries, there is not a single landlocked Arab country. North Africa provides a particularly dramatic illustration of this lost potential for development. Stretching from Egypt to Morocco, North Africa is blessed with thousands of kilometres of coastline. Its proximity to Europe and Africa makes it one of the choicest locations for other emerging markets. Everywhere else in the world, direct access to the sea translates into lower transport costs and better prospects for manufacturing. However, the Middle East defies the economic laws of gravity: it has coastal access without market access.

Another aspect of geography, where the Middle East is hugely favoured but is failing to materialise its inherent advantage, is urbanisation. At least 50% of the total population in the MENA region (excluding Yemen) resides in urban areas. Latest measures of urban concentration place the region ahead of other developing countries, including those in Latin America, making it one of the most urbanised regions of the world. Recent evidence suggests that urbanisation can deliver concrete benefits to firms: by locating in urban centres, firms enjoy not only proximity to markets but superior access to a range of mutually supportive activities (skills, machinery, suppliers, resources and the like). These agglomeration economies are simply absent in the Middle East.

Even if Arab economies do not suffer from the kind of structural geographic barriers that hinder prosperity in Africa, the region's divided equilibrium is supported by both political and geo-political imperatives. As discussed above, the region's pervasive trade barriers serve an important political function through a political economy of protectionism, where closed economic borders are part of a broader political strategy for regime survival. The rentier characteristics of Arab economies also engender a sense of autonomy from integration, since external windfalls from oil and aid have insulated the region from pressures for economic cooperation. Economic fragmentation is also underpinned by a geopolitical equilibrium that relies on the divide and rule strategies of regional and global hegemons. In short, the absence of a strong domestic constituency, internal rivalries and dependence on external powers have frustrated past attempts at regional economic integration.

It is true that part of this failure to integrate regional markets is rooted in production structures that look more similar than different, but an enabling policy response – through a regionally coordinated industrial policy, for instance – has also been lacking. Dismantling regional trade barriers has been an economically desirable but politically inexpedient step. The demographic and political shifts in the region call for a new logic of economic integration. Given the multiple costs of fragmentation outlined above, a vibrant private sector requires soft borders and thick markets. In this milieu, fostering regional economic cooperation is arguably the single most important collective action problem that the region has faced since the fall of the Ottoman Empire. The Arab world will ultimately need not just political commons, but also regional economic commons, that could serve as incubators for entrepreneurship, employment and growth.

See Also

▶ Business Politics in the Gulf;
▶ Labour Markets in the Arab World;
▶ Oil and Politics in the Gulf: Kuwait and Qatar;
▶ Rent Seeking

Bibliography

Assad, R. 2014. Making sense of Arab labour markets: The enduring legacy of dualism. *IZA Journal of Labour and Development* 3(6). doi:10.1186/2193-9020-3-6.

Cammett, M., and I. Diwan. 2014. Conclusion: The political economy of the Arab uprisings. In *A political economy of the Middle East*, 3rd ed., ed. A. Richard and J. Waterbury. Boulder: Westview Press.

Cammett, M., and N. Silti. 2014. Perceptions of public welfare and political mobilization in the Middle East: Preliminary evidence from Egypt and Tunisia. Paper presented for the workshop on 'The Pulse of the Arab Street', Université Paris-Dauphine, 11–12 October 2014.

Diwan, I. 2013. Understanding revolution in the Middle East: The central role of the middle class. *Middle East Development Journal* 5(1): 30.

Diwan, I. 2014. Fifty years of fiscal policy in the MENA region. *Unpublished draft*, Economic Research Forum, Cairo.

Diwan, I., P. Keefer, and M. Schiffbauer. 2013. The effect of crony capitalism on private sector growth in Egypt. *Working paper*. Harvard Kennedy School, Cambridge, MA.

Hertog, S., M. Valeri, and G. Luciani. 2013. *Business politics in the Middle East*. London: Hurst Publishers.

Malik, A. 2014. A requiem for the Arab development model. *Journal of International Affairs* 68(1): 95–115.

Malik, A., and B. Awadallah. 2013. The economics of the Arab Spring. *World Development* 45: 296–313.

Rijkers, B., C. Freund, and A. Nucifora. 2014. All in the family: state capture in Tunisia. *World Bank Policy Research Working* Paper 6810.

Saudi Gazette. 2014. 11 October.

Shui, L., and P. Walkenhorst. 2010. Regional integration: status, developments, and challenges. In *Trade competitiveness in Middle East and North Africa: Policies for export diversification*, ed. J.R. López-Cálix, P. Walkenhorst, and D. Ndiameé. Washington, DC: World Bank.

Political Economy of Institutional Change: Illustrations from the Ottoman Empire

Metin M. Coşgel

Abstract

This article sheds light on the political economy of the Ottoman Empire through the lens of its policy on tax collection and technology adoption. Like all rulers, the Ottomans were constrained in their abilities to implement economic policies as they wished. In addition to having limited resources and technology, they faced political constraints that altered the feasibility, desirability and outcomes of economic policies. In taxation, they allowed the tax bases and rate structures to vary significantly across regions to balance revenue maximisation with political power. In technology, despite adopting advancements in military technology immediately, they waited almost three centuries to fully sanction the printing press because it would have undermined the ability of religious authorities to confer legitimacy.

Keywords

Conquest; Constraints; Fiscal regime; Legitimacy; Loyalty; Political economy; Power; Religion; State; Taxation

JEL Classifications

D8; H2; J4; L3; M5; N4

Political incentives are crucial to explain economic policy choices. This article shows how political forces shaped economic policies and outcomes in the Ottoman Empire by focusing on its systems of taxation and technology adoption. Although conventional caricatures of the Ottoman Empire have sometimes painted it as a stagnant monolith, recent studies have shown that the rulers of the Empire accommodated both change and the status quo, depending on their political incentives (Coşgel 2015; Pamuk 2012).

Like all rulers, the Ottomans would have preferred to govern as they wished, but they faced numerous political constraints imposed by the reaction of the general public to their policies or by the interests of powerful organised groups, such as the nobility, the military or religious authorities, who sought to maximise their own welfare, even at the expense of others. The rulers did not have unlimited control over economic outcomes, because they depended on the general public for revenue and they drew power from organised groups who could lose their ability to provide legitimacy or revolt if their interests were sufficiently threatened. Because of the conflicts of interests, the rulers were ultimately constrained in their abilities to tax the population and regulate the economy.

For a simplified model of the political economy of an empire, consider a society that consists of a ruler, the general population and an organised group that acts as an intermediary between the ruler and the population and whose role it is to support (legitimise) the ruler. The people produce a surplus, part of which can be extracted by the ruler for his own consumption. The objective of the ruler is to maximise his consumption. The organised group, such as the nobility or the legal or military authority, has a choice between supporting the ruler or inciting a revolt against him. If it chooses to legitimise the ruler, the ruler can extract a surplus from the population in the form of tax payments. The group's support raises the size of the surplus by making the people view the ruler as legitimate and pay taxes without resistance. In return, the ruler shares his surplus with the group to elicit its support. Alternatively, rather than support the ruler, the group can choose to incite a revolt against him. If the revolt succeeds, the group would obtain the surplus, but if it fails, the ruler gets the surplus and the group gets nothing.

In the Ottoman Empire, the Sultan could acquire legitimacy mainly from religious authorities (*şeyhülislam* and the *ulamā*), nobility (*a'yān*) and military authorities (the *sipāhī* and janissary organisations). Religious authorities could confer legitimacy through loyalty, promoting the belief that the Sultan had the right to rule and the power to provide protection and other public goods and services – and that he should therefore have the right to collect taxes. Their power depended on their role in the transmission of knowledge, an essentially oral process in early Ottoman society prior to the introduction of the printing press. Secular authorities could also confer legitimacy through loyalty, based on their powers as eminent individuals who led tribes, owned land or other resources, belonged to prominent families, and mediated people's relationship with the state in their capacities as regional representatives, tax collectors and managers of public order and civil disputes. By contrast, military authorities could confer legitimacy through force, based on their comparative advantage in using manpower and weapons. Military commanders were in principle at the Sultan's disposal, ready to employ their troops to secure his legitimacy.

We now use this simplified setting to analyse how political forces affected economic choices in the Ottoman Empire, and more specifically in their policies on taxation and technology regulation. The argument, in a nutshell, is that in newly conquered lands the Ottomans implemented a tax system by considering not just the efficiency of rates, bases and collection methods, but their effect on the chances of establishing legitimacy. In regulating technology they adopted some of the new technologies, but banned others in a way that considered their effects on not just productivity but also on the abilities of organised groups to legitimise the ruler. We detail below how the legitimising relationships and conflicts of interest between the rulers, tax collectors and taxpaying public affected the system of taxation in newly conquered districts and how the ruler's relationship with the religious, secular and military authorities affected the introduction of new technologies.

Taxation

Starting from a small tribe settled in northwestern Anatolia at the end of the 13th century, the Ottomans kept expanding in the next three centuries and eventually built a vast Empire that spanned the area from the Black Sea in the north to Egypt and the Arabian Peninsula in the south, and from the Persian Gulf in the east to central Europe and North Africa in the west. Conquering land from multiple predecessor states, they inherited the tax systems of various legal and political traditions that needed to be moulded into a coherent whole and applied to local conditions. The Ottomans developed a tax system that reflected various regional idiosyncrasies from the customs and administrative practices of preceding states, indicating that some things were harder to change than others and that political economy constraints played an important role in shaping the final outcome.

To see regional idiosyncrasies of the tax bases and rate structures, consider the variation in personal taxes. Under the conventional system observed in Anatolia, personal taxes were based on adult males, and the tax rate varied by marital status and land ownership. The subjects in Hungary, on the other hand, paid personal taxes in terms of the gate (*kapı*) tax, for which the unit of taxation was the household, rather than adult males, and the tax amount did not change by marital status or land ownership. Moreover, personal taxes were not even fully implemented in all areas (though non-Muslim subjects throughout the empire paid a poll tax called *cizye*). In Jerusalem and surrounding districts, for example, the Ottomans did not introduce the *çift* tax or any other form of personal tax systematically levied on individuals or households. Trade and production taxes also varied a great detail among regions (Coşgel 2005, 2015).

With each new conquest, the Ottomans thus faced a basic choice between preserving the existing system of taxation in newly conquered lands or changing it to conform to the system prevailing in other parts of the empire. Although efficiency considerations (e.g. cost of collection, incentive implications) affected the choice, political constraints were also important.

Political constraints on taxation can be categorised into two broad groups: those originating from the reactions of the general population and others coming from legitimising agents who collected taxes on behalf of the government. To start with the first group, note that the general population naturally resents taxation and prefers stable, secure incomes, indicating that they would oppose (or even revolt against) the rates being raised significantly or the system being changed drastically. Opposition to tax policies has been one of the most common reasons for popular uprisings in history, most evident during conquests. Unless changing the tax system clearly eliminated excessively oppressive elements of previous taxes (as may have been the case for labour services), the general population likely preferred the status quo over change, for fear that change could mean higher taxes and worse conditions. They could flee the land or revolt against the new regime if the changes were perceived to be too burdensome. Even if the Ottomans discovered an existing tax system to be inefficient, they had to carefully weigh their desire to change it for efficiency gains against the rising likelihood of political instability and revolt against their regime. An inefficient tax system could survive if political constraints prevented a ruler from changing it.

The Ottomans were not free to change the tax codes as they wished because of political realities surrounding conquest, assimilation and stability. Even if they could have increased the tax revenues in Hungary by changing personal taxes from being based on the household as a whole to a differential rate structure based on the characteristics of its individual members, they would have met stiff resistance from those who would have paid higher taxes. Because of this resistance, they could not have implemented the change easily. Once the tax code of a region was adopted, changing it would have been difficult because the general population, accustomed to paying taxes under a familiar system, and powerful groups with vested interests in this system would have continued to resist the change and initiate a revolt against the Ottomans.

The second group of political constraints on taxation were related to the system of tax collection, the way the Ottomans appointed agents to collect taxes and allocated tax revenues among these agents. Organised in a multi-tiered system, the Ottoman government consisted of multiple hierarchical levels that divided the Empire into provinces, the provinces into districts and the districts into fiefs or other administrative units. To support offices at lower levels, the central government assigned some of the tax revenues directly to governors of provinces (variously denoted in the registers as *mīr mīrān*, *paşa*, *beylerbeyi*), district officials (*mīr liwā*, *sancakbeyi*), and holders of small and large fiefs (*tımār* and *za'āma*). In the resulting system of tax collection and revenue allocation, you could have one village paying taxes to the central government, their neighbours in the next village paying them to the provincial government and still others paying them to the district government or a local fiefholder (cavalrymen or military commander).

There were also tribal leaders who somehow possessed the right to collect the tax revenues of some villages and landholders, who similarly held the rights to collect taxes privately (*mülk*) or jointly with the government (under a system called *mālikāne dīvānī*). Typically, these were rights the Ottomans had preserved from the system that they inherited upon conquest or assigned through negotiations with powerholders.

The Ottomans allocated tax collection rights among their agents in part for economic reasons, such as to accommodate differential abilities of agents to assume risks (due to the variability of the tax base) or to measure the tax base. According to a quantitative analysis of tax assignment in the Ottoman Empire, revenues allocated to local government officials included a higher proportion of variable taxes than those allocated to the provincial and central treasury, indicating that the variance of the tax base affected the allocation (Coşgel and Miceli 2005). Economic factors also influenced the government's decision on which contractual form (rent, wage, share) to adopt in employing agents for tax collection.

For a complete explanation of how the Ottomans allocated tax revenues among private landholders, tribal leaders and other agents, we need to go beyond purely economic factors and consider how tax collectors affected the ruler's legitimacy. In general, tax collectors not only acted as government officials in their districts, but also as legitimising agents that could enhance the ruler's ability to extract the surplus. This ability gained even greater significance in newly conquered lands. Acting as local representatives of the new ruler, carefully chosen tax collectors could solidify his legitimacy and raise the share of the gross surplus that he could extract from the general population for his own consumption. In addition, using local elites for tax collection could significantly expand elite-coalition by serving as an important commitment device for ensuring their support for the centralised ruler.

Tax collectors could affect the legitimacy of the Ottoman sultans through both force and loyalty during this period (Coşgel and Miceli 2009). If a newly conquered population included powerful individuals who could be bribed into using their power for tax collection, the Ottomans could be better off relying on these individuals for the service than military officials appointed from the centre. In the same vein, if the local population included individuals with leadership qualities that could generate loyalty by encouraging the citizens to accept the Ottoman ruler's right to rule and his ability to provide protection and other public goods and services, the Ottomans could receive legitimacy more effectively through them than an Ottoman official appointed from the centre but unknown in this region. If the local leaders in newly conquered areas thus had superior ability to provide legitimacy through force or loyalty, these qualities could supersede economic considerations in the assignment of tax revenues, and the Ottomans could be better off using them as agents to raise the share of the surplus that the public would pay voluntarily as taxes.

By including political constraints in the analysis, we offer a more complete explanation of why the Ottomans appointed tribal leaders and private individuals as tax collectors in some areas. They preserved the rights of some landholders to collect taxes after conquest under the *mālikāne dīvānī* system in eastern Anatolia because in those regions these leaders had a comparative advantage in force and loyalty that was essential to collect taxes on behalf of the Ottomans, a right that was granted in exchange for a share of the tax revenue. The Ottomans similarly appointed tribal leaders in some regions as tax collectors so that the Ottoman rule would be established within the institutional constraints of conquest politics. They assigned some of the tax revenues to Bedouin tribes in the Fertile Crescent, for example, so as to establish the Ottoman rule in the desert frontiers. Although purely economic concerns of expanding into new territories might have suggested to the Ottomans that they should replace previous systems of taxation and collection with more efficient schemes that could be adopted from other parts of the empire, political constraints sometimes required them to work within the parameters of existing orders and capitalise on the comparative advantages of local agents who could better legitimise their regime through force and loyalty.

Technological Change

The Ottomans showed a mixed reaction to developments in technology. While readily adopting some of the new technologies, they rejected others outright or delayed their adoption for a long time. They paid close attention to advances in military technology (e.g. gunpowder, firearms and cannons) and assimilated them into the army and the navy swiftly. In adopting the printing press, by contrast, they took nearly three centuries after the invention of moveable type to sanction and offer explicit support for printing in Ottoman Turkish (in Arabic characters).

The delayed adoption of the printing press has attracted significant attention in the literature because of its implications for Muslim attitudes towards science and technology and for the indication that this may have contributed to the economic underdevelopment in the Islamic world. The generalist literature and Eurocentric approaches have typically offered *ad hoc* religious and cultural explanations of the attitude towards the printing press, such as religious conservatism towards Western technology and the inability of Muslim culture and institutions to keep up with changing times.

The problem with these approaches is their failure to explain why the rulers banned the printing press. For a complete explanation of this type, one would have to make controversial *ad hoc* assumptions about not just the conservative values of the general public but also about the motivations of Ottoman rulers in banning the new technology. This could be consistent with traditional analyses of government that viewed rulers as benevolent protectors who served the interests of the general public. If the printing press was expected to harm everyone, the ruler could be justified to ban it. Given the social and economic benefits of mass printing to the general public, however, this approach makes little sense and falls short of offering a complete explanation for why a ruler would reject a useful technology that had the potential to raise productivity. It does not seem justified to presume that the rulers and all organised groups had the same values and interests as the general public.

Contrary to traditional analyses of government, recent political economy models have made standard economic assumptions about the motivations of rulers. In this view, all members of the society, including the rulers, seek to maximise their own interests, even at the expense of the rest of the society. A society's reaction to a new technology would thus reflect not necessarily a unified social, cultural or religious concern, but an outcome of the strategic interaction of all affected parties.

In general, according to the political economy framework developed above, a society would adopt a new technology depending on its effect on the size of the surplus available to the ruler for taxation, the ability of religious and secular authorities to legitimise the ruler and the probability of a successful revolt. If the rulers expected new developments to raise the revenue available to them while having a positive effect on legitimacy or revolt, they would be eager to adopt them swiftly. But if the adoption of a new technology was likely to increase the probability of a successful revolt, despite its positive effects on productivity and output, they could oppose it if the net effect to them was negative. In the same vein, the general public and organised groups would be expected to favour technologies that improved their welfare and to oppose others. The final outcome would depend on a multitude of factors, including the direction and magnitude of how a new technology was likely to affect the welfare of the general population, the amount of the surplus available to the ruler, the probability of revolt and the abilities of organised groups to legitimise the ruler and the remuneration that the ruler was willing to pay for their services.

One of the possible outcomes was the quick adoption of a new technology. As noted, the Ottomans were usually eager to adopt new developments in military technology. Realising the advantages of gunpowder weapons, they integrated them into their army as swiftly as possible. They not only kept pace with developments in gunpowder, firearms and cannons, but displayed ingenious organisational skills by pioneering the establishment of a permanent standing army (the Janissaries) specialised in the use of these weapons, well before the European powers. They showed such remarkable success in assimilating gunpowder technology in their army and navy that by the mid-15th century they had achieved a clear logistical and firepower superiority over their European and Asian adversaries.

The Ottomans were generally eager to accept a new military technology because they expected it to raise the revenue available to them without significant adverse consequences to their basis for legitimacy or the probability of a successful revolt against their rule. A new military technology could raise the size of the surplus available to the Ottomans by expanding their revenue base through conquests and tributes or by helping them protect existing revenues from being confiscated by adversaries equipped with the new technology. Advances in military technology raised the ability of military authorities to legitimise the ruler without affecting the probability of a successful revolt significantly because the Ottomans took various measures to maintain their monopoly in organised violence. They controlled rural banditry by striking bargains with their leaders and incorporated bandits into the system by recruiting them as irregular soldiers. They also implemented a system of periodic rotation of offices through which government officials were rotated on a more or less regular schedule and were prevented from forming potential alliances with rebellious movements (Coşgel et al. 2013). Because advances in military technology could raise the size of the surplus available to them without having a significant effect on the legitimisation relationship between the rulers and the legal community or on the probability of revolt, the rulers accepted them eagerly.

Another possibility was an indifferent attitude towards a new technology. Two scenarios could lead to this outcome. First, the cost or benefit of a new technology on the ruler's surplus, his legitimacy or the probability of revolt could be too small to catch his attention or to provoke great enthusiasm or opposition. The introduction of various consumption goods invented elsewhere, such as eyeglasses and clocks, was often greeted with indifference, and consequently they were adopted by default. Similarly, despite their opposition to the printing press in the 15th century, the Ottomans were previously indifferent to the adoption of paper, a Chinese invention. Although these were great scholarly and scientific accomplishments, the cost and benefit to the ruler's surplus, his legitimacy and probability of revolt were often negligible, and there was no major reason for them to occupy the ruler's attention extensively. The Ottoman rulers typically did not show great enthusiasm (regarding their effect on their concerns) toward this type of development; nor did they oppose them vehemently.

Another scenario leading to indifference was when a new development could provide substantial benefits to one of the ruler's concerns (surplus, legitimacy or revolt), but was also accompanied by a similarly substantial cost to another concern. If, for example, a new technology was expected to raise the ruler's surplus significantly, but only at the cost of raising the likelihood of revolt by a comparable magnitude, the rulers could react to this technology with indifference. Despite expecting substantial benefits, they could be in no rush to adopt it quickly; nor would they have a great need to reject or delay its adoption, because the net effect could be negligible. Other concerns would determine the outcome.

While some developments were eagerly accepted and others faced lukewarm reception, still others were rejected outright. The preceding discussion has shown that just because a new development was likely to reduce the surplus available to the ruler did not mean it would be rejected. Despite causing a significant loss in surplus, it could still be adopted if the loss was offset by benefits in legitimacy and/or the likelihood of revolt. But if the ruler suffered a loss not just in surplus but also in legitimacy and likelihood of revolt, or if the loss in one of these areas was substantially greater than benefits in others, he could reject a new development immediately or delay its adoption. Some of the new developments in science, technology and institutions were rejected or regulated carefully because the net loss would have been too large.

As noted, a good example of this was the immediate opposition and significant delay in the adoption of the printing press. Within decades after the appearance of Gutenberg's first book published by moveable type in Germany, the Ottoman sultan is said to have issued an edict that banned printing in Ottoman Turkish (in Arabic characters) in 1485. Although the authenticity of this edit has been questioned, the evidence nevertheless indicates that some type of severe restriction on printing in Ottoman Turkish was in place. Despite clear awareness of the new printing technology and successful reproduction of it within Ottoman lands by religious minorities, the process of accepting the printing press was extremely slow. Even after the rulers started to relax the ban in 1726, they continued to regulate the operation closely by granting permission only to selected individuals, prohibiting publication on religious subjects and appointing a committee of scholars to review and proofread contents for accuracy. It was not until well into the 19th century that mass printing technologies became the convention and new techniques were adopted quickly and used commonly. The lithographic press was adopted within a few decades after its invention, and a press was established to print an official newspaper.

The Ottomans did not delay adopting the printing press because they were unaware of its invention or lacked the technical expertise. What makes the heavy regulation of the printing press puzzling is that its adoption could have raised the society's taxable surplus. Although the effect of printing on economic welfare would have been less in the Ottoman Empire than in Europe (because of differences in wage and literacy rates), it is reasonable to expect the effect on economic activities (indirectly on tax revenues) to be positive and substantial, directly through its effect on the market for books (e.g. the Quran and other religious texts) and indirectly through its effect on human capital and positive externalities that would have benefited other sectors. Judging solely by its effect on economic activities, Ottoman rulers would have been better off adopting the printing press immediately.

According to the political economy approach, the rulers were unenthusiastic about the printing press from its invention until the 19th century because they expected it might undermine the ability of religious authorities to confer legitimacy and might increase the likelihood of a successful revolt against their reign, even though it could raise the size of the surplus available to them. In early modern Muslim societies, the religious authorities had a monopoly on providing legitimacy through indoctrination because the transmission of knowledge depended on oral technology, and the authorities had a vast comparative advantage in this type of transmission. The introduction of the printing press would have altered this ability by changing the technology of transmitting knowledge and diminishing the comparative advantage of religious authorities. The general public could obtain knowledge directly from books or from literate individuals not necessarily affiliated with religious authorities. The rulers probably feared mass printing also because of its potential effect on successful revolt. Mass printing could be a very effective weapon in inciting rebellion, as was the case in the American Revolution and the Protestant Reformation. Although the printing press could have raised the surplus available to the rulers by a margin, Ottoman rulers still chose to ban it because it would have jeopardised their legitimacy and increased the likelihood of a successful revolt (Coşgel et al. 2012).

The examples on tax and technology adoption suggest that the institutional milieu in Ottoman Empire was not defined by an outright absence of change. Since changes in taxation or technology affected not just the productivity of the population but also the legitimising relationships between the rulers and their agents, institutional change was only permitted as long as it was consistent with ruler's incentives. In deciding whether to change the tax bases and rate structures in newly conquered territories and whether to adopt newly invented technologies, the Ottomans had to weigh economic concerns against political incentives carefully.

All of this highlights the primacy of politics.

See Also

- ▶ History and Comparative Development
- ▶ Labour Markets in the Arab World
- ▶ Political Economy of Arab Uprisings
- ▶ Religion and Economic Development
- ▶ Taxation of Wealth
- ▶ Technical Change
- ▶ Technology

Bibliography

Coşgel, M.M. 2005. Efficiency and continuity in public finance: The Ottoman system of taxation. *International Journal of Middle East Studies* 37(4): 567–586.

Coşgel, M.M. 2015. The fiscal regime of an expanding state: Political economy of Ottoman taxation. In *Fiscal regimes and the political economy of premodern states*, ed. A. Monson and W. Scheidel. Cambridge: Cambridge University Press.

Coşgel, M.M., and T.J. Miceli. 2005. Risk, transaction costs, and government finance: The distribution of tax revenue in the Ottoman Empire. *Journal of Economic History* 65(3): 806–821.

Coşgel, M.M., and T.J. Miceli. 2009. State and religion. *Journal of Comparative Economics* 37: 402–416.

Coşgel, M.M., T.J. Miceli, and J. Rubin. 2012. The political economy of mass printing: Legitimacy, revolt and technological change in the Ottoman Empire. *Journal of Comparative Economics* 40: 357–371.

Coşgel, M.M., B. Ergene, H. Etkes, and T.J. Miceli. 2013. Crime and punishment in Ottoman times: Corruption and fines. *Journal of Interdisciplinary History* XLIII (3): 353–376.

Pamuk, S. 2012. Political power and institutional change: Lessons from the Middle East. *Economic History of Developing Regions* 27(s1): S41–S56.

Political Economy of Unearned Foreign Income

An Application for Non-oil Producing Muslim States

Faisal Z. Ahmed
Princeton University, Princeton NJ, USA

Abstract

This entry argues that foreign aid and remittances constitute a form of "unearned foreign income" that has affected the public finances and shaped political outcomes in the non-oil producing Muslim countries in North Africa, the Middle East, and South Asia. Aid and remittance flows have stabilized authoritarian rule in this "broader" Middle East and North Africa (MENA) region by reducing the likelihood of conflict, fostering corruption, and extending the duration of non-democratic governments.

Keywords

Foreign aid; Remittances; Unearned income; Governance; Non-democracy

JEL

F3; P16; F24; F35

Introduction

Unearned foreign income comprised of foreign aid and workers' remittances is an important source of income for many developing countries. Increasingly, scholars are evaluating the macro-*political* consequences of these capital inflows. One strand of this nascent scholarship conceptualizes aid and remittances inflows as a form of "unearned income" with non-tax like properties for government finances. In doing so, this strand of research links the political economy ramifications of aid and remittances to the politically pernicious effects associated with rentier states (e.g., authoritarianism, corruption, political violence). Building on this conceptualization, this entry argues that aid and remittance inflows have entrenched authoritarian governance and rule in many non-oil producing Muslim countries in North Africa, the Middle East, and South Asia (a region, I will call "broader MENA.")

Unearned Foreign Income and Authoritarian Governance

Aid and Remittances as Unearned Foreign Income

Unearned Income Unearned income is a concept in economics that has different meanings and implications depending on the theoretical framework used. Political economists broadly view unearned income as non-tax government revenue (e.g., Madhavy 1970; Besley and Persson 2010). For them, the distinction between non-tax and tax revenue has implications for political development: governments that derive a greater share of their revenues from taxation tend to be more democratic and more accountable to their populations (e.g., Tilly 1992).

Foreign Aid as Unearned Foreign Income Foreign aid is the international transfer of capital, goods, or service from a country or international organization ("donor") for the benefit of the recipient country (i.e., its government, its population). These transfers can involve financial resources, technical advice and training, or commodities (e.g., food or military equipment). In many instances, the resources can take the form of grants or concessional credits (e.g., export credits). In practice, the most common type of foreign aid is official development assistance

(ODA), which aims to promote development and combat poverty. The primary source of ODA is bilateral grants from one country's government to another, and while a small fraction of aid may "bypass" a recipient government to a non-government organization, the vast majority of aid directly enters a recipient government's revenue. Since this revenue is foreign and not derived from domestic taxation, it fits the definition of unearned income.

Remittances as Unearned Foreign Income Whereas foreign aid is a transfer of resources between governments, remittances do not directly enter a government's revenue base. A remittance is a transfer of money by a foreign worker to an individual in his or her home country. Frequently, this transfer occurs between family members. Measuring remittance flows can be arduous since they may flow through unofficial channels and many recipient (developing) countries often lack the capacity to accurately track, record, and tax these capital inflows (Chami et al. 2008). However, remittances may enter a government's revenue base indirectly. For instance, via an expenditure-switching mechanism, a government may reduce its provision of certain welfare goods (e.g., health care or education), forcing migrant households to purchase them instead. This mechanism frees resources for governments to spend elsewhere, such as on the military and government salaries (Abdih et al. 2012; Ahmed 2012).

Unearned Income and Authoritarian Governance

Unearned Income and Governance The ability to "collect" revenue is central to the viability of any state (Levi 1988); without it, the state cannot carry out its basic functions such as providing security. States can collect their revenue from taxation (e.g., from individuals, firms, land, capital, consumption, international trade) and non-tax sources (e.g., revenues from the state's production of natural resources, foreign aid). On the latter, the availability of unearned income as a viable and important source of government revenue (especially since the nationalization of oil production in many countries during the 1970s) has sparked an active research agenda that theorizes and empirically evaluates the effects of unearned income on politics, such as democracy, institutions, and civil war (Ross 2013).

Scholars of the Middle East were the first to articulate the relationship between unearned income and democracy, while investigating whether the prevalence of "rents" contributed to authoritarian governance in that region (Mahdavy 1970). The central argument in the rentier state literature is that governments funded by external rents are freed from the need to raise taxes, which makes them less accountable to their citizens and to less committed to democratic governance. Many qualitative and quantitative studies have empirically substantiated this relationship (see Ross 2013 for an overview).

This negative relationship between unearned income and democratic governance may depend on the quality of preexisting institutions. For instance, Tornell and Lane (1999) advance a model showing how, upon receiving a positive shock to fiscal capacity (e.g., a resource boom, foreign aid), a state with weak institutions may suffer from a "voracity effect" in which powerful groups compete for and squander the windfall, in addition to diminishing any pro-growth effects. Focusing on the political ramifications, Robinson et al. (2006) formulate a parallel model demonstrating that when institutions are weak ex-ante, increases in unearned income lead to excessive public employment and patronage (e.g., corruption). In these weak institutional settings, unearned income can finance additional forms of patronage goods such as the co-option of political rivals, religious leaders, business elites, and the military. The latter can strengthen a government's repressive capacity to fend off revolutionary threats (Bueno de Mesquita et al. 2010).

Unearned Foreign Income and Patronage in Non-democracies Patronage politics is particularly salient for political survival and stability in non-democracies (Bueno de Mesquita et al. 2003). As a consequence, when compared to governments in democracies, governments in autocracies will place greater weight on expenditures for patronage in their spending decisions. Building on this framework, Ahmed (2012) develops a game theoretical model in which foreign aid and remittances expand a government's revenue base (akin to traditional forms of unearned income) and finances patronage in non-democratic regimes. Unearned foreign income does so via two distinct channels: an income effect and a substitution effect.

As a transfer of resources between governments, aid directly increases government revenue. And since governments in autocracies tend to allocate a greater share of their revenues towards patronage, a higher share of their aid is spent towards patronage goods. This process constitutes an "income effect."

In contrast, remittances do not directly enter a government's revenue base. Rather, the model identifies a substitution effect. Households will spend a portion of their remittance income on certain substitutable welfare goods (e.g., better quality health care and/or education for their children). The government "observes" these expenditures and marginally decreases its own provision of that welfare good, diverting those unused funds towards patronage. In the model's equilibrium, autocratic governments have a greater incentive (i.e., higher utility) to divert resources towards patronage from higher levels of remittance income. Thus, the substitution effect is magnified in non-democratic regimes. In combination, Ahmed shows that higher levels of aid and remittances finance more patronage in non-democracies. Thus, aid and remittances should stabilize governments in non-democracies, for example, by lowering the likelihood of institutional change and government turnover.

Unearned Foreign Income in the Broader MENA Region

Within this theoretical framework, the next two sections present empirical evidence linking unearned foreign income to non-democratic governance in the non-oil producing Muslim countries in North Africa, the Middle East, and South Asia; a region I will call "broader" Middle East and North Africa (MENA).

Unearned Foreign Income in Broader MENA

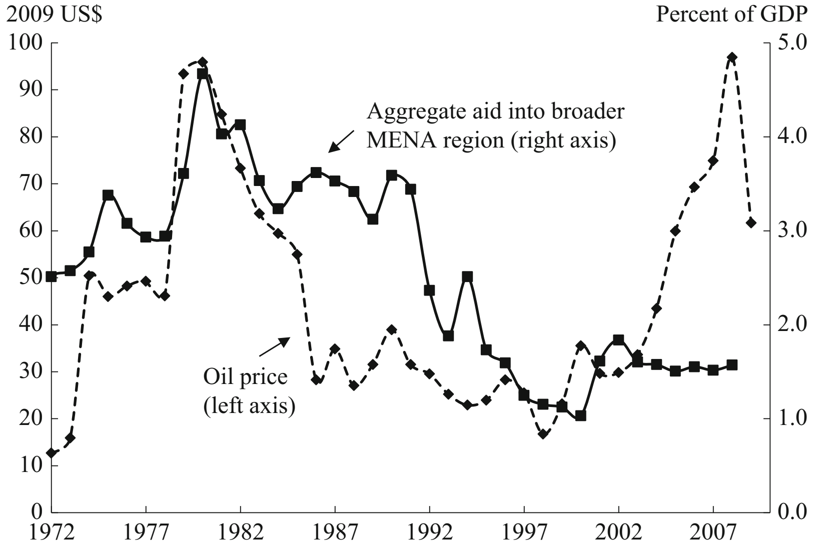
Since the 1970s, countries in the broader MENA region have received large sums of unearned foreign income; a substantial portion of which originates from oil producing countries in the Persian Gulf (e.g., Saudi Arabia, Kuwait, the United Arab Emirates). And, oil price booms and busts explain a large share of the temporal variation in these capital flows.

Gulf oil producers ramped up their aid disbursements following the first oil price shock in 1973. Between 1974 and 1994, Gulf donors doled out 1.5% of their GDP, which amounted to 13.5% of all aid given out over this period – overwhelmingly, this aid favoured non-oil producing Muslim countries (Neumayer 2003). Certainly, some of the motivation for this assistance was political, as the Gulf countries were at the same time trying to subdue unrest stemming from the huge inequality among their co-religionists (between the oil haves and have-notes), and seeking to "assure them[selves] a clear position of dominance within the Muslim world" (Kepel 2002, 69–70). Much of this Gulf aid looked like unearned income (e.g., oil rents) flowing to the state, as it came primarily in the form of block grants to recipient governments with comparatively few strings attached (Hunter 1984).

The oil boom in the 1970s also affected remittance inflows into the broader MENA region. In the aftermath of the 1973 oil crisis, labour from different countries in North Africa, South Asia, and Middle East migrated in great numbers to the oil-exporting countries in the Middle East. The first wave of workers (totaling about 500,000) migrated from non-oil producing Gulf States, such as Jordan, Palestine, and Yemen. In the latter part of the decade, Gulf States began to recruit a large number of South Asian workers from India, Pakistan, and Bangladesh. For example, it is estimated that the number of Pakistani workers jumped from roughly 500,000 in 1975 to over 1.25 million in 1979. By the early 1980s, there may have been some 3.5–4.65 million migrants, in a combined labour force of 9–10.2 million workers (Choucri 1986). This large movement of labour generated large capital flows in the form of workers' remittances from Gulf oil producers to a variety of non-oil producing labour exporting countries in the Middle East (e.g., Jordan), Africa (e.g., Mali), and South Asia (e.g., Pakistan).

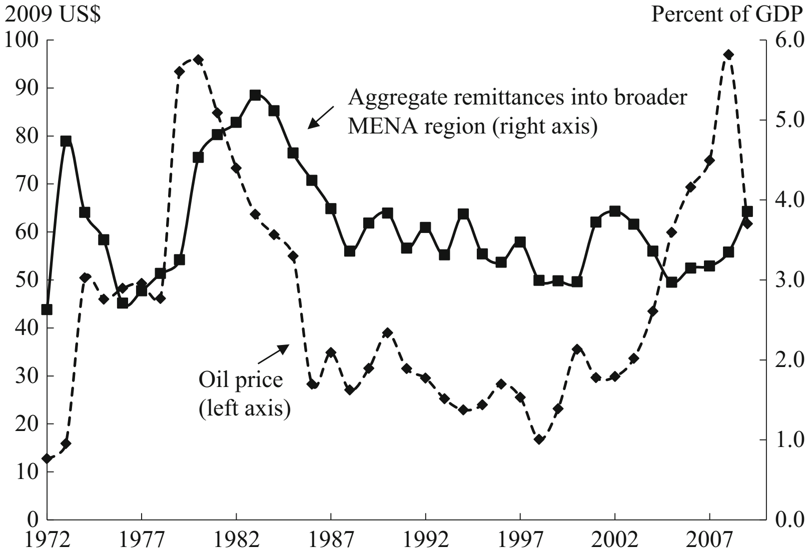
Temporal Variation

Inflows of aid and remittances into the broader MENA region have varied over time and tracked the world price of oil. Figures 1 and 2 depict movements in the price of oil (right axis) and superimpose the *aggregate* annual amount of aid (Fig. 1) and

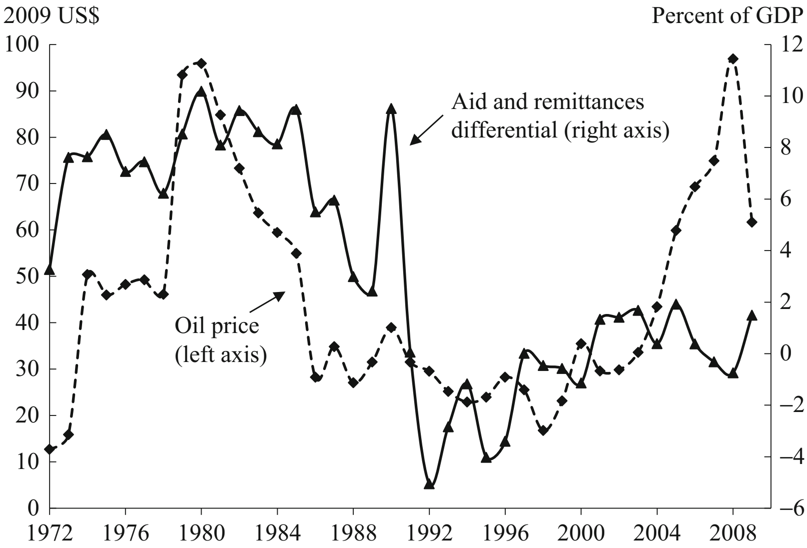


Notes: Foreign aid includes disbursements from all donors (as reported by the OECD) into the broader MENA region. This includes disbursements from both OPEC and non-OPEC (e.g., the United States, the UK) donors. Data from the World Bank World Development Indicators and the author's calculations

Political Economy of Unearned Foreign Income, Fig. 1 Aggregate foreign aid in the "broader MENA" region



Political Economy of Unearned Foreign Income, Fig. 2 Aggregate remittances in the "broader MENA" region (*Notes*: Remittances include flows from all countries (as reported by the World Bank) into the broader MENA region. This includes remittances from the Persian Gulf and elsewhere (e.g., Western Europe and North America). Data from the World Bank World Development Indicators and the author's calculations)



Political Economy of Unearned Foreign Income, Fig. 3 Difference in unearned foreign income between non-oil producing Muslim and non-Muslim recipients (*Notes*: Data from the World Bank World Development Indicators and the author's calculations)

remittances (Fig. 2) into the broader MENA region (measured as a share of the region's GDP). In the early 1970s, aggregate aid and remittance inflows were quite low but then increased sharply following the first oil shock in 1973. These inflows remained high through about 1984 and then declined after 1985 as oil prices tanked. As oil prices recovered in the 1990s, aid and remittance flows began to rise again. In the 2000s, the correlation (positive) between oil prices and these capital flows (in particular aid) tends to weaken. On balance, Figs. 1 and 2 suggest a strong correlation between oil price and aggregate flows of unearned foreign income into the broader MENA region.

Interestingly, the recipients of oil price-induced aid and remittance flows were primarily non-oil producing Muslim countries. Conversely, similar "shocks" resulting from variation in oil prices did not affect aid and remittances to non-oil producing *non*-Muslim nations. Figure 3 provides such evidence. It plots movements in oil prices (left axis) and superimposes the difference in average of foreign aid and remittance inflows between non-oil producing Muslim and non-oil producing non-Muslim countries (right axis). A positive differential implies the "typical" Muslim countries received higher amounts of aid and remittances (as a share of GDP) relative to the typical non-oil producing non-Muslim recipient.

Figure 3 shows that between 1970 and 2000, this "differential" in aid and remittances tracks the price of oil. As oil prices rose in the 1970s, so did the differential. Moreover, it remained large and positive during the decade of high oil prices (1974–1984). Countries in the broader MENA region tended to receive substantial inflows of aid and remittances (i.e., between 6% and 10% of additional GDP per annum) than non-Muslim recipients. As oil prices plummeted after 1986, so did the differential. On average, Muslim and non-Muslim countries tended to receive similar amounts of aid and remittances (as a share of their respective country GDPs).

Evidence

Establishing Causality

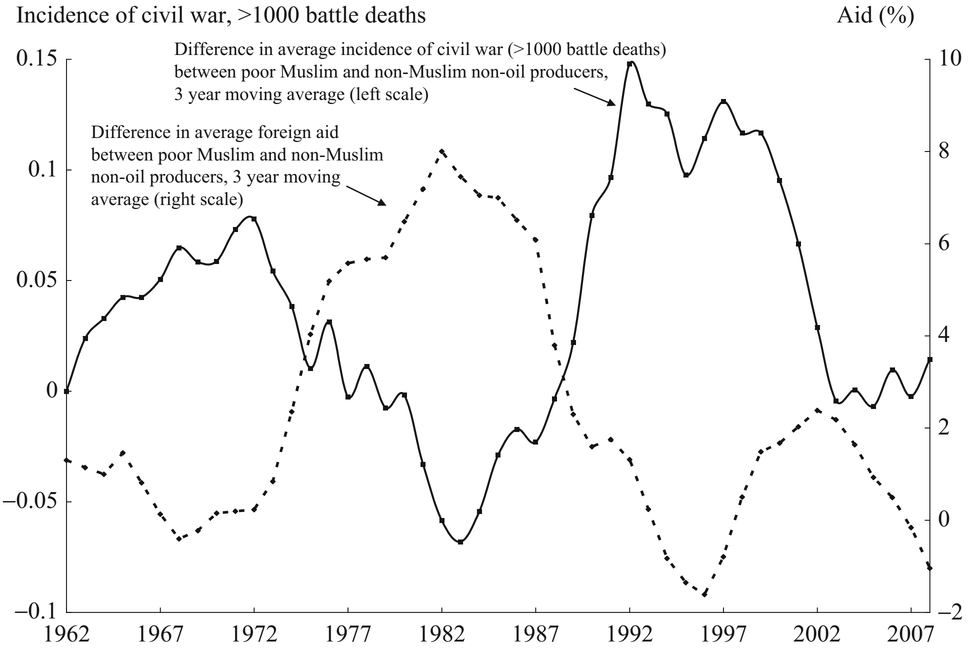
Evaluating the causal effect of unearned foreign income on political outcomes is problematic as these income flows are plausibly endogenous with the "politics" in the receiving states. Gauging a causal effect therefore requires some plausibly exogenous source of variation in these income flows that is uncorrelated with underlying political and economic conditions in the receiving states. Figures 1 and 2 show that foreign aid and remittances inflows into the broader MENA region are correlated with world oil prices, and these prices are plausibly exogenous to political and economic conditions in poor, non-oil producing Muslim countries (e.g., Jordan, Bangladesh). Thus, variation in oil prices can serve as a plausibly exogenous source of variation in aid and remittance flows to the broader MENA region. Moreover, Fig. 3 shows that in comparison to non-Muslim countries, Muslim countries were the main beneficiaries of the oil-price induced aid and remittance "shocks." Indeed, several studies leverage these facts to evaluate the causal effect of aid and remittances on various political outcomes in the broader MENA region.

Foreign Aid and Civil War

Ahmed and Werker (2015) evaluate whether foreign aid affects political stability. They argue that aid can "buy" political stability (often by strengthening a state's repressive capacity and autocratic institutions), while declines in aid foster instability. They measure instability using the incidence of civil war. To identify the causal impact of aid on conflict, they leverage a difference-in-differences (DID) strategy to show that periods of higher aid receipts (due to higher oil prices) are associated with a lower probability of conflict in the broader MENA region.

Figure 4 captures their core findings. The figure examines the relationship between the aid differential and the conflict differential across Muslim and non-Muslim countries. Muslim countries in the broader MENA region experienced less conflict than non-Muslim countries when they received comparatively more aid. As the aid differential reversed due to lower oil prices from the mid-1980s until the early 2000s, Muslim countries were substantially more likely to experience civil war.

The differential effects of aid on political violence are large and statistically significant. Ahmed and Werker (2015, Table 1) present DID estimates in which the aid windfall between 1973 and1985 made countries in the broader MENA region 7 percentage points more stable (i.e., less conflict prone) compared to non-Muslim aid recipients. The end of the windfall fostered a relative rise in political violence as Muslim countries became 11 percentage points more likely to be engaged



Political Economy of Unearned Foreign Income, Fig. 4 Foreign aid and civil war (Source: Fig. 5 in Ahmed and Werker (2015))

Political Economy of Unearned Foreign Income, Table 1 The impact of remittances on authoritarian governance

Dependent variable	Corruption		Transfers	Salaries
			(% Gov. expenditures)	
Method of estimation	2SLS	2SLS	2SLS	2SLS
	(1)	(2)	(3)	(4)
Remittances (% GDP)	0.323	−0.143	−3.248	3.39
	(0.151)**	(0.066)**	(1.258)**	(0.876)***
Autocracy		−11.798		
		(0.178)*		
Remit. × Autocracy		1.361		
		(0.618)**		
No. obs	863	863	305	315

Notes: Results from Ahmed (2013), Tables 6, 8, and 11. Estimation via 2SLS. Robust standard errors clustered by government are reported in parentheses. *, **, *** = Significant at 10, 5, 1%, respectively. All specifications control for GDP per capita (% annual), log GDP per capita (1995 US$), log population, POLITY autocracy score, year trend, and country fixed effects. Remittances are instrumented with p(oil)*distance to Mecca. In columns 1 and 2, the dependent variable is ICRG corruption index (range: 1–6) where a higher value implies greater corruption. In column 3, the dependent variable is government transfers and subsidies (% government expenditures). In column 4, the dependent variable is government compensation of employees (% government expenditures)

in civil war, while non-Muslim countries became slightly more stable (around 1 percentage point).

Remittances and Corruption

Autocrats frequently permit corruption as a strategy of political survival. It can be a means to reward loyal supporters and erode the provision of public goods in favour of private government goods, such as patronage (Bueno de Mesquita et al. 2003). Thus, given the viability of corruption as a strategy to maintain political stability in autocracies, Ahmed (2013) investigates whether or not remittances may be a conduit for corruption. To that effect, Ahmed interacts oil prices with a Muslim country's distance to Mecca to construct a cross-national and time-varying instrumental variable for remittances received by countries in the broader MENA region. To hone in on patronage-based government corruption, Ahmed uses the ICRG corruption index, which measures "actual or potential corruption in the form of excessive patronage, nepotism, job reservations, 'favor-for-favors', secret party funding, and suspiciously close ties between politics and business."

Leveraging this research design, the paper establishes two key results. First, remittances foster corruption, particularly in countries with preexisting authoritarian politics. Second, remittances do so by allowing governments to divert spending from welfare goods to patronage. Table 1 reports a snapshot of the econometric analysis from Ahmed (2013), highlighting these main findings. The coefficient estimate in column 1 implies that a 3-percentage point increase in aggregate remittances raises the corruption index by about 1 index point. Column 2 evaluates whether a country's preexisting quality of authoritarian politics magnifies the effect of remittances on corruption. The positive coefficient on the interaction effect (Remit. x Autocracy) implies that remittances received in countries with more autocratic politics has a *greater* causal impact on corruption.

The remaining columns in Table 1 show that a substitution effect is a plausible channel through which remittances foster corruption. Columns 3 and 4 show that remittances cause governments in the MENA region to decrease their expenditures on welfare goods (column 3) and shift those to patronage in the form of greater compensation for government employees (column 4). The latter is a common measure of patronage in developing countries, as it frequently reflects the government's incentives to channel spending to targeted constituencies (Keefer 2007). Remittances are also negatively associated with government expenditures on health care and education (Ahmed 2013, Table 11).

Unearned Income and Political Stability

Finally, in combination, foreign aid and remittances have stabilized governments in the broader MENA region. Table 2 below reports the core results from Ahmed (2012). That paper employs an instrumental variables strategy to show that greater inflows of unearned foreign income in the broader MENA region caused governments to experience a lower likelihood of losing office between 1973 and 2004. The coefficient estimate in column 1 implies that a 1 percentage increase in unearned foreign income (relative to GDP) lowers the likelihood that a government will fall out of power by about 5 percentage points (in that calendar year). Moreover, this stabilizing effect is larger for governments in countries with stronger authoritarian political institutions (column 2). Since many of the governments in the MENA region were authoritarian over the sample period, the results in columns 1 and 2 imply that unearned foreign income increased the duration of autocrats in non-oil producing Muslim countries. King Hussein in Jordan, as well as the military dictatorships in Bangladesh and Pakistan during the 1970s and 1980s, illustrates this claim.

Column 3 presents evidence of both an income effect and a substitution effect associated with unearned foreign income inflows. In this specification, the dependent variable is a government's expenditures on transfers and subsidies as a share of total government expenditures. A negative regression coefficient implies that the variable shifts the allocation of government spending away from the provision of welfare goods to another type of government spending. Given this interpretation, the negative coefficient on aid and remittances implies that increases in unearned foreign income reduce a government's share of expenditures on welfare goods. This observation is consistent with a substitution effect. Of course, as Ahmed (2012) explicitly models, an autocrat may want to spend some fraction of its aid on welfare goods. The positive coefficient on aid is consistent with this income effect.

Conclusion

How governments derive their revenues can have important political ramifications. The availability of unearned income or non-tax revenue, in particular, can make governments less accountable to their populations and fund various strategies of political survival (e.g., patronage, repression). Governments in these states tend to be non-democratic. Building on this framework, this entry argues and presents evidence that foreign aid and remittances have constituted a form of unearned foreign income that has fostered authoritarian politics in countries in the broader MENA region.

Political Economy of Unearned Foreign Income, Table 2 Unearned foreign income and authoritarian stability

Dependent variable	Government turnover		Transfers (% gov exp.)
Method of estimation	2SLS	2SLS	2SLS
	(1)	(2)	(3)
Aid and remit. (% GDP)	−0.046		−1.509
	(0.022)**		(0.785)*
Aid and remit. × Autocracy		−0.294	
		(0.135)**	
Aid (% GDP)			1.363
			(0.777)*
No. obs	1639	1639	315

Notes: Results from Ahmed (2012), Tables 4 and 6. Estimation via 2SLS. Robust standard errors clustered by government in parentheses. *, **, *** = Significant at 10, 5, 1%, respectively. In columns 1 and 2, the dependent variable is equal to 1 if the government loses office in that year, and zero otherwise. In columns 1 and 2, the specifications control for log GDP per capita (1995 US$), GDP per capita growth (% annual), duration splines, and indicator variables for finite-term, incidence of low and high internal discontent, country and year fixed effects. Splines are duration time, duration time squared, and duration time cubed. Column 3 controls for log GDP per capita (1995$). These coefficients and a constant are not reported. Aid and remittances are instrumented with the p(oil) × Muslim country dummy

These findings offer broader insights on the political economy of unearned income. First, since a large share of the aid and remittances emanated from oil producing countries in the Persian Gulf (and was driven primarily by movements in oil prices), it stands to reason that the pernicious political and economic effects of the "resource curse" can be exported abroad via capital outflows (Ahmed et al. 2016). Second, since non-oil producing Muslim countries comprised the unique recipients of a shock in unearned foreign income in the 1970s (relative to non-oil non-Muslim countries), the processes described in this entry offer a plausible explanation for the "democratic deficit" that has emerged in Islamic states (Huntington 1993). Finally, while this entry has focused on the political ramifications of aid and remittances in the broader MENA region, there is no reason to assume that these effects manifest themselves only in this region (e.g., Ahmed (2017); Doyle 2015). Thus, a fruitful area of future research will assess whether or not unearned foreign income generates similar effects in other regions with different underlying political structures.

See Also

- Development Economics
- Dutch Disease and Foreign Aid
- Foreign Aid

Bibliography

Abdih, Yasser, Ralph Chami, Jihad Dagher, and Peter Montiel. 2012. Remittances and institutions: Are remittances a curse? *World Development* 40 (4): 657–666.

Ahmed, Faisal Z. 2012. The perils of unearned foreign income: Aid, remittances, and government survival. *The American Political Science Review* 106 (1): 146–165.

Ahmed, Faisal Z. 2013. Remittances deteriorate governance. *The Review of Economics and Statistics* 95 (4): 1166–1182.

Ahmed, Faisal Z. 2017. Remittances and incumbency: Theory and evidence. *Economics and Politics* 29 (1): 22–47.

Ahmed, Faisal Z., and Eric D. Werker. 2015. Aid and the rise and fall of conflict in the Muslim world. *Quarterly Journal of Political Science* 10 (2): 155–186.

Ahmed, Faisal Z., Daniel Schwab, and Eric D. Werker. 2016. *The political transfer problem*, Working paper.

Alesina, Alberto, and Beatrice Weder. 2002. Do corrupt governments receive less foreign aid? *American Economic Review* 92 (4): 1126–1137.

Bardhan, Pranab. 1997. Corruption and development: A review of issues. *Journal of Economic Literature* 35 (3): 1320–1346.

Besley, Timothy, and Torsten Persson. 2010. State capacity, conflict, and development. *Econometrica* 78 (1): 1–34.

Blattman, Christopher, and Edward Miguel. 2010. Civil war. *Journal of Economic Literature* 48 (1): 3–57.

Chami, Ralph, Adolfo Barajas, Thomas Cosimano, Connel Fullenkamp, Michael Gapen, and Peter Montiel. 2008. *The macroeconomic consequences of remittances*. Washington, DC: International Monetary Fund.

Choucri, Nazli. 1986. The hidden economy: A new view of remittances in the Arab World. *World Development* 14 (6): 697–712.

de Mesquita, Bueno, Alastair Smith Bruce, Randolph M. Siverson, and James D. Morrow. 2003. *The logic of political survival*. Cambridge, MA: MIT Press.

de Mesquita, Bueno, Alastair Smith Bruce, Randolph M. Siverson, and James D. Morrow. 2010. Leader survival, revolutions, and the nature of government finance. *American Journal of Political Science* 54 (4): 936–950.

Doyle, David. 2015. Remittances and social spending. *The American Political Science Review* 109 (4): 785–802.

Hunter, Shireen. 1984. *OPEC and the third world: The politics of aid*. London: Croom Helm.

Huntington, Samuel P. 1993. The clash of civilizations? *Foreign Affairs* 72 (3): 22–49.

Keefer, Phillip. 2007. Clientelism, credibility, and the policy choices of young democracies. *American Journal of Political Science* 51 (4): 804–821.

Kepel, Gilles. 2002. *Jihad: The trail of political Islam*. Trans. Anthony F. Roberts. Cambridge, MA: Harvard University Press.

Levi, Margaret. 1988. *On revenue and rule*. Berkeley: University of California Press.

Mahdavy, Hussein. 1970. The patterns and problems of economic development in Rentier States: The case of Iran. In *Studies in economic history of the middle east*, ed. M.A. Cook, 428–467. London: Oxford University Press.

Neumayer, Eric. 2003. What factors determine the allocation of aid by Arab countries and multilateral agencies ? *Journal of Development Studies* 39 (4): 134–147.

Robinson, James A., Ragnar Torvik, and Thierry Verdier. 2006. Political foundations of the resource curse. *Journal of Development Economics* 79 (2): 447–468.

Ross, Michael L. 2001. Does oil hinder democracy? *World Politics* 53 (3): 325–361.

Ross, Michael L. 2013. *The politics of the resource curse: A review*, Working paper.

Tilly, Charles. 1992. *Coercion, capital, and European states, AD 990–1992*. Cambridge, MA: Blackwell.

Tornell, Aaron, and Philip R. Lane. 1999. The voracity effect. *American Economic Review* 89 (1): 22–46.

Werker, Eric D., Faisal Z. Ahmed, and Charles Cohen. 2009. How is aid spent?: Evidence from a natural experiment. *American Economic Journal: Macroeconomics* 1 (2): 225–244.

World Bank. 2010. *World development indicators*. Washington, DC: World Bank.

Political Institutions, Economic Approaches To

Timothy Besley and Torsten Persson

Abstract

Political institutions affect the rules of the game in which politics is played. Economists now have theoretical approaches to explain the impact of institutions on policy, and empirical evidence to support the relevance of the theory. This article sketches a framework to inform discussions about how political institutions shape policy outcomes. It does so using four examples: majoritarian versus proportional elections; parliamentary versus presidential government; whether to impose term-limits on office holders; and the choice between direct and representative democracy. Each example illustrates how theory and data can be brought together to investigate a specific issue.

Keywords

Accountability; Adverse selection; Agency; Citizen initiative and referendum; Coalition government; Corruption; Direct democracy; Majoritarian elections; Moral hazard; Parliamentary government; Plurality rule; Political competition; Political institutions, economic approaches to; Presidential government; Proportional representation; Rent seeking; Representative democracy; Representation; Separation of powers; Spatial voting models; Targeted public spending; Term limits; Treatment effects

JEL Classifications

D72

Introduction

Political institutions play a key role in shaping economic policies. Economists now have theoretical approaches to explain this claim and empirical evidence to support it. Political institutions affect the rules of the game in which politics is played. For the most part the term ‘institutions’ is taken to mean formal rules as embodied in constitutions, and other forms of legislation. However, it may also refer to norms and informal rules.

Two basic categories of political institutions are electoral rules and forms of government. The former term refers to features such as district magnitudes and electoral formulas that translate votes into seats. It also refers to the rules for selecting candidates and for governing their tenure in office. The latter category refers to such questions as whether the systems is presidential or parliamentary, how decision making powers are divided between central and local governments or between executive and legislature, and whether citizens have a direct say in policymaking via referenda.

Our aim in this article is to sketch an intellectual framework that informs discussions about how political institutions may shape policy outcomes. We do this by way of specific examples, referring to recent research on the topic – we do not attempt to provide a comprehensive overview of theoretical modelling or empirical knowledge. In each case, the example illustrates the potential for theoretical frameworks to shape thinking on the topic backed up with empirical analysis.

When political scientists debate democratic institutions, they frequently use two metrics for their performance – accountability and representation. The former refers to the way in which political institutions make politicians (and to some degree bureaucrats) answerable for their actions. The second refers to whether the policies and/or policymakers fairly reflect the population as a whole.

Translated into the language of economics, these two performance dimensions correspond well to two main conflicts of interest that arise in representative democracies – those between politicians and citizens and those between groups of citizens with competing economic interests. Accountability deals predominantly with the former and representation with the latter. As normative criteria, the welfare underpinnings of these metrics are somewhat vague, but they do provide a useful way of thinking about the positive effects of political institutions.

Economic models for studying accountability are mostly based on some form of *agency* approach. Such models assume that there exist problems of hidden actions (moral hazard) and hidden types (adverse selection) in politics. Politicians typically have career concerns which lead them to seek re-election. Voters decide whether or not to re-elect based on the record of politicians. To make the problem interesting, there has to be some conflict of interest between politicians and voters. The simplest (and most widely used model) supposes that this is due to opportunities for rent seeking (or effort avoidance) among politicians. The question is then how much of this conflict of interest rubs off on to policy choice in equilibrium, that is, when voters and politicians are behaving rationally and optimally. There is now a large body of literature using such models. Political institutions can affect policy in such models in three main ways: affecting the information that voters have to assess politician performance, directly affecting incentives of politicians to extract rents, and affecting the kinds of people of who are selected for public office (See Besley 2006, for a broad survey of agency models and their uses).

Economic models for studying representation rely on some kind or another of a spatial framework. These models envisage citizens being located at different points in the space according to their underlying economics interests (such as their age or ability) and their social interests (such as ethnicity). The classic Downsian model of political competition (Downs 1957) falls in this class and many subsequent developments have built on its insights. More recent work has tried to make the framework more tractable by supposing that voting is probabilistic – there is a random element in the ballots cast by voters, and politicians can therefore not be exactly sure how policies translate into voting outcomes. In standard models, competition is directly over policies without regard to who is being asked to carry these policies out. More recent approaches have looked at the problem of picking policymakers to deliver these policies. This is particularly important when modelling the credibility of policies being offered (See Persson and Tabellini 2000, for a broad survey of spatial models of policymaking and their uses).

In this article, we illustrate the main themes of the recent literature by focusing on four examples of how political institutions shape policy. Two of these examples deal with electoral rules, two with forms of government, broadly defined. Two of the examples are motivated mainly from cross-country empirical applications, while the other two are motivated more from studies of within-country variation (Persson and Tabellini 2003, discuss empirical work on cross-country studies of political institutions, while Besley and Case 2003, survey within-country (cross-state) studies for the United States). Thus, Sections 2–5 discuss, in turn, the policy consequences of adopting proportional or plurality elections, the effects of parliamentary or presidential forms of government, the consequences of term limits for elected politicians, and the impact of direct or representative democracy. Section 6 concludes.

Proportional or Majoritarian Elections

Political scientists often describe a key trade-off in electoral systems: electoral formulas based on plurality rule promote accountability at the expense of representation, while formulas based on proportional representation (PR) errs on the other side of the trade-off. Recent theoretical work by economists has analysed the consequences for governments spending of having legislative seats awarded by plurality rules rather than PR – an issue closely related to representation. The key idea is relatively straightforward (see Persson and Tabellini 1999; Lizzeri and Persico 2001; Milesi-Feretti et al. 2002). If candidates with the highest vote shares win every seat at stake in a district, rather than seats in proportion to their vote shares, it becomes more attractive to target spending to small and geographically concentrated groups of voters (The same will hold true if each district has small magnitude, that is, represents a small share of the electorate). This tilts equilibrium policy towards spending programmes with benefits targeted to particular geographical groups, not the electorate at large, and (perhaps) towards higher overall spending.

Empirical work has sought to evaluate these predictions using cross-national data. Long-term inertia in the broad features of electoral systems makes it necessary to rely on the cross-sectional variation in the data, which, together with the nonrandom selection of electoral systems, raises a number of statistical issues. These issues are tackled by a variety of methods in Persson and Tabellini (2003, 2004), who classify actual electoral systems according to their electoral formula (classifying by district magnitude gives similar results) and approximate geographically non-targeted spending by welfare-state programmes, such as pensions and unemployment insurance. Their results indicate that a reform from an all-PR to an all-plurality-rule system would cut welfare spending by about two per cent of GDP in the long run. Such an electoral reform would cut overall government spending by a substantial five per cent of GDP.

The underlying theory works off the incentives of politicians and takes party structure as given. Yet it is a well documented fact that PR promotes a more fractionalized party system than plurality rule (see, for example, Lijphart 1990). Austen-Smith (2000) studies a model where redistributive tax policy is set in post-election bargaining, assuming that the number of parties is, exogenously, higher under PR than plurality rule. He shows that this produces higher taxes and spending under PR. Bawn and Rosenbluth (2006) and Persson et al. (2005) obtain a similar prediction but endogenize the number of parties. In their models of parliamentary democracy, they show that coalition governments spend more than single-party governments under each electoral rule. We should still observe higher spending in PR systems, but this is an indirect effect of a larger number of parties increasing the incidence of coalition government. Persson et al. (2005) derive an empirical way of discriminating between the indirect effect and the direct effect via the incentives of politicians. Using panel data for parliamentary democracies since 1960, they find that the higher overall spending observed under PR is entirely due to its more fractionalized party systems and hence more frequent coalition governments than under plurality rule.

A second body of theory relates to the accountability of politicians under alternative electoral systems. The key idea here is that extraction of rents – or, more generally, corruption – is better deterred the more swiftly the probability of re-election responds to performance (see Myerson 1993; Persson and Tabellini 2000). Large district magnitude achieves this by allowing easier entry and a larger number of candidates than small districts. Personal ballots impose individual accountability and stronger incentives than party-list ballots, which impose only collective accountability. In other words, systems where a larger number of lawmakers are elected in each district, and systems where they are elected on personal rather than party-list ballots, are both expected to reduce rent extraction by politicians. Empirically, Persson and Tabellini (2003) find quite sizeable effects in the hypothesized direction on different perception indexes of corruption, or on inefficiency in the delivery of government services.

Presidential or Parliamentary Government

How well voters can hold politicians accountable also depends on the form of government. This insight goes far back in political writing. For example, James Madison insightfully discussed various aspects of the separation of powers in his contributions to *The Federalist Papers*. Economists have recently produced modern versions of the argument as to how separation of powers across political offices may serve to limit conflicts of interest between voters and their elected representatives. Extending the agency model of Ferejohn (1986), Persson et al. (1997) show that separating the proposal powers over taxes and spending creates a conflict between politicians that enables voters to better discipline their power to extract rents when in office.

This approach is extended to include issues of representation by Persson et al. (2000), who analyse how different forms of government shape fiscal policy by embedding different forms of legislative bargaining in spatial voting models. They assume that presidential systems have a more extensive separation of powers across legislators than parliamentary systems. On the other hand, as in Huber (1996) and Diermeier and Feddersen (1998), parliamentary systems make the government subject to a confidence requirement of the legislature, whereas a presidential system does not (the president is directly elected). These two institutional features shape the legislative bargaining, such that legislative majorities in presidential systems become less stable than in parliamentary regimes. If majorities re-form, issue by issue, different minorities are pitted against each other for different issues on the legislative agenda. As a result, broad spending programmes suffer at the expense of targeted spending. Moreover, the lack of a stable legislative majority means that there is no well-defined residual claimant on government revenue. This reduces the incentives to boost overall taxation and spending. Overall, we should thus expect presidential regimes to be associated with lower total spending and smaller broad (non-targeted) spending programmes than parliamentary regimes.

Persson and Tabellini (2003, 2004) confront these predictions with data, in which real-world forms of government are classified as parliamentary or presidential, depending on whether the executive is subject to the continual confidence of the legislature. For broad welfare state programmes, they find the hypothesized result only among long established democracies, among which presidential regimes spend less, by about two per cent of GDP. For overall spending the results are very robust across samples and in line with the basic hypothesis. Whether the results are obtained by OLS, instrumental variables or matching methods, the finding is that presidential regimes have smaller governments by at least five per cent of GDP – again, a large number.

Term Limits or No Term Limits

Political accountability is achieved in part by re-election chances responding to performance while in office. This resembles the kind of contractual relations that arise in a market context and provide workers with incentives. However, the relationships between politicians and voters are not contractual – they resemble something closer to a fiduciary relationship. While political parties may have a role in disciplining politicians, the ultimate sanction is an electoral one: poorly performing incumbents are removed from office by the voters.

The frequency of re-election and the number of terms that a politician can serve become important institutional choices in shaping electoral accountability. The agency model of politics referred to above provides a tool to approach these issues. The theory suggests two ways of thinking about term limits: incentive effects and selection effects. Incentive effects arise because politicians who face a shorter time horizon are less obliged to please voters. Whether this increases or reduces the quality of policy is moot. On the one hand, politicians facing term limits may have less incentive to please voters and hence may follow their private agendas. But they may also pander to voters, eschewing hard decisions that impose short-run costs in exchange for long-run benefits. This latter effect can lead term-limited politicians to act more in the voters' interests. Either way, if electoral incentives matter, then we should expect term limits to shape political decisions. Terms limits will also induce a selection effect. Politicians have to be elected to lame-duck terms. Rational voters should anticipate this when deciding whether to (re)elect them, which will make politicians elected to lame-duck terms better than average. Such positive selection may counteract any adverse incentive effect.

US states provide a natural experiment for looking at the impact of term limits, because governors are subject to such limits in around half the states. This allows two kinds of comparisons: across time – governors when they are up against a term limit versus their first (non-term limited) period in office – and across states – term- limited versus non-term-limited governors.

Besley and Case (1995) identify the effect of a term limit from the difference between first and second terms in office for incumbents facing term limits. Controlling for state fixed effects and year effects, and using annual data from the 48 continental US states from 1950 to 1986, they find that a variety of policy measures are affected by term limits. Specifically, state taxes and spending are higher in the second term when term limits bind in states that have them. Such limits tend to induce a fiscal cycle, with states having lower taxes and spending in the first gubernatorial term than in the second. More recently, List and Sturm (2006) have applied these ideas to environmental policies at the US state level and also find evidence of a term-limit effect. They observe that the way in which environmental interests are represented in policy may depend on whether the governor is in his last term in office.

Term limits have also been advocated as solutions to institutional distortions in legislatures. A good example is the committee system in the US Congress, which puts a premium on seniority of politicians and thus, effectively, a lower performance threshold for incumbents with a resulting diminution in accountability (see Dick and Lott 1993, for development of this argument).

A host of studies look for effects of announced retirements on voting behaviour in Congress. On the whole, it has been difficult to find evidence of a last-period effect. For example, Lott and Bronars (1993) analyse Congressional voting data from 1975 to 1990 and find no significant change in voting patterns in a representative's last term in office. McArthur and Marks (1988) look at Congressional behaviour in a lame-duck session of Congress: in post-election sessions, members who have not been re-elected are at times called upon to vote on legislation before the swearing in of the new Congress. They find that lame-duck representatives were significantly more likely in 1982 to vote against automobile domestic content legislation than were returning members.

Direct or Representative Democracy

Whether polities should use some element of direct democracy as part of their political institutions is widely debated. The two most famous examples are US states and Swiss cantons, which display considerable variation in their reliance on citizen initiatives and referenda. From a theoretical point of view, issues of accountability and representation are important in thinking through these issues.

Some commentators (for example, Denzau et al. 1981) emphasize the role of initiatives in reducing rent-seeking by government and hence enhancing accountability in the political process. This underpins a number of studies investigating whether jurisdictions that permit initiatives have smaller governments. For example, Matsusaka (1995) regresses government expenditures and revenues on a number of control variables for a panel of 49 US states (Alaska excluded) sampled over a 30-year period at five-year intervals from 1960 to 1990. He includes year effects, but not state fixed effects, since the presence of initiatives is largely fixed within states over time. His main finding is a strong negative effect on expenditures of access to the initiative. Matsusaka (1995) also finds some evidence that the effect is strongest where the number of citizen signatures required for a referendum is low. Similarly, Pommerehne (1990) shows that Swiss cantons using the initiative indeed have smaller governments.

Others emphasize the fact that initiatives can change the representation of policy preferences. A large body of empirical evidence from political science supports the lack of congruence of policy and voter preferences on a variety of issues (see Besley and Coate 2000, for references).

Gerber (1999) considers how, given a set of policy preferences in a legislature, the availability of the initiative could change the equilibrium policy bargain. Moreover, the legislature may make such a change pre-emptively, that is, it is sufficient for legislators to anticipate the *possibility* of an initiative at a later date. Hence, the possibility of initiatives forces a greater agreement between voter preferences and policy outcomes, on the assumption that representatives elected to the legislature have views that are out of step with the citizens as large. Similar conclusions follow from the theoretical analysis of Besley and Coate (2000) but for quite different reasons. They develop a model in which initiatives affect electoral outcomes. They argue initiatives have an impact via *issue unbundling*. In general elections, many issues are decided at once, which may result in non-salient issues being distorted away from the preference of a majority. Initiatives allow such issues to be unbundled from other issues in the election. Besley and Coate show that this can change the probability distribution of a range of policy outcomes and the composition of candidates who are chosen to run. Both of these theoretical approaches, as well as many popular discussions of initiatives, imply that citizen initiatives are a device for bringing policy into line with public opinion.

One strand of empirical literature on initiatives has used data from US states to test whether public opinion and policy outcomes are closer together in initiative states. For example, Lascher et al. (1996) and Camobreco (1998) investigate whether the link between aggregate measures of policy outcomes and public opinion is closer when states allow citizens' initiatives. They find no significant effect. With respect to specific policy issues, Gerber (1999) uses cross-sectional state variation from the 1990s and compares stances on an array of policies. She finds significant differences (at the ten per cent level) for personal income taxes (initiative states lower); highway, natural resources and hospital spending (initiative states higher in all cases); and the implementation of three-strike legislation (initiative states lower). Gerber looks in greater detail at the death penalty and parental consent laws for abortion, using public opinion data to estimate median voter preferences. With cross-sectional data for 1990, she runs a logistic regression that interacts whether a state has an initiative with public opinion, and finds that states with initiatives mirror public opinion on abortion and the death penalty more closely, even though these policies are not directly determined via initiatives.

Final Remarks

The examples discussed above illustrate how knowledge in the field has benefited from research targeted towards understanding specific issues, even though these issues can be nested in broader debates about accountability and representation. Theoretical and empirical research on the boundary between economics and political science has uncovered systematic relationships between political institutions and policy outcomes, and is currently being extended to new domains of economic policymaking.

One challenge for the future is to study what determines changes in institutions over time. It is evident that studying how political institutions work, the focus of the discussion here, is a necessary part of research on institutional change. From a theoretical point of view, it is important to understand whose interests are served by particular institutional arrangements and how policies change as a consequence of them. For practical purposes, this will likely be a piecemeal agenda dealing with specific constitutional arrangements rather than examining constitution design from the ground up. This is why the kind of nuts and bolts issues illustrated in our four examples provides the basis for further progress in the field.

Much of the empirical research, so far, has adopted a relatively simple approach, in which political institutions are taken as given and the hypothesized institutional impact is the same across political, social and economic conditions. As is well known from the microeconometric treatment literature, this can easily lead to biased estimates. Current research has started to address non-random selection of political institutions as well as the likely existence of heterogenous treatment effects, where the effect of a specific institutional reform depends on social and historical preconditions. Measurement and econometric testing of these complex issues would benefit greatly from new theoretical research on the endogeneity and conditional effects of institutional reform.

See Also

- ► Political Competition

We are grateful to Jenny Mansbridge for helpful comments.

Bibliography

Austen-Smith, D. 2000. Redistributing income under proportional representation. *Journal of Political Economy* 108: 1235–1269.

Bawn, K., and F. Rosenbluth. 2006. Short versus long coalitions: Electoral accountability and the size of the public sector. *American Journal of Political Science* 50: 251–265.

Besley, T. 2006. *Principled agents: The political economy of good government*. Oxford: Oxford University Press.

Besley, T., and A. Case. 1995. Does political accountability affect economic policy choices? Evidence from gubernatorial term limits. *Quarterly Journal of Economics* 110: 769–798.

Besley, T., and A. Case. 2003. Political institutions and policy choices: Evidence from the United States. *Journal of Economic Literature* 41: 7–73.

Besley, T., and S. Coate. 2000. *Issue unbundling via citizens' initiatives*, Working paper no. 8036. Cambridge, MA: NBER.

Camobreco, J.F. 1998. Preferences, fiscal policies and the initiative process. *Journal of Politics* 60: 819–829.

Denzau, A.T., R.J. Mackay, and C. Weaver. 1981. On the initiative-referendum option and the control of monopoly government. In *Tax and expenditure limitations*, ed. H.F. Ladd and T.N. Tideman. Washington, DC: The Urban Institute.

Dick, A.R., and L. Lott Jr. 1993. Reconciling voters' behavior with legislative term limits. *Journal of Public Economics* 50: 1–14.

Diermeier, D., and T. Feddersen. 1998. Cohesion in legislatures and the vote of confidence procedure. *American Political Science Review* 92: 611–621.

Downs, A. 1957. *An economic theory of democracy*. New York: Harper.

Ferejohn, J. 1986. Incumbent performance and electoral control. *Public Choice* 50: 5–25.

Gerber, E. 1999. *The populist paradox: Interest group influence and the promise of direct legislation*. Princeton: Princeton University Press.

Huber, J. 1996. The vote of confidence procedure in parliamentary democracies. *American Political Science Review* 90: 269–282.

Lascher Jr., E.L., M.G. Hagen, and S.A. Rochlin. 1996. Gun behind the door? Ballot initiatives, state policies and public opinion. *Journal of Politics* 58: 760–795.

Lijphart, A. 1990. The political consequences of electoral laws, 1945–1985. *American Political Science Review* 84: 481–496.

List, J., and D. Sturm. 2006. How elections matter: Theory and evidence from environmental policy. *Quarterly Journal of Economics* 121: 1249–1281.

Lizzeri, A., and N. Persico. 2001. The provision of public goods under alternative electoral incentives. *American Economic Review* 91: 225–245.

Lott Jr., J.R., and S.G. Bronars. 1993. Time series evidence on shirking in the U.S. House of Representatives. *Public Choice* 76: 125–149.

Matsusaka, J.G. 1995. Fiscal effects of voter initiative: Evidence from the last 30 years. *Journal of Political Economy* 103: 587–623.

McArthur, J., and S.V. Marks. 1988. Constituent interest vs. legislator ideology: The role of political opportunity cost. *Economic Inquiry* 26: 461–470.

Milesi-Feretti, G.-M., R. Perotti, and M. Rostagno. 2002. Electoral systems and the composition of government spending. *Quarterly Journal of Economics* 117: 609–657.

Myerson, R. 1993. Effectiveness of electoral systems for reducing government corruption: A game theoretic analysis. *Games and Economic Behavior* 5: 118–132.

Persson, T., G. Roland, and G. Tabellini. 1997. Separation of powers and political accountability. *Quarterly Journal of Economics* 112: 1163–1202.

Persson, T., G. Roland, and G. Tabellini. 2000. Comparative politics and public finance. *Journal of Political Economy* 108: 1121–1161.

Persson, T., G. Roland, and G. Tabellini. 2005. *Electoral rules and government spending*. Mimeo: Stockholm University.

Persson, T., and G. Tabellini. 1999. The size and scope of government: Comparative politics with rational politicians. 1998. Alfred Marshall lecture. *European Economic Review* 43: 699–735.

Persson, T., and G. Tabellini. 2000. *Political economics: Explaining economic policy*. Cambridge, MA: MIT Press.

Persson, T., and G. Tabellini. 2003. *The economic effect of constitutions*. Cambridge, MA: MIT Press.

Persson, T., and G. Tabellini. 2004. Constitutional rules and fiscal policy outcomes. *American Economic Review* 94: 25–64.

Pommerehne, W.W. 1990. The empirical relevance of comparative institutional analysis. *European Economic Review* 34: 458–469.

Politics and Economics

V. K. Borooah

No great powers of persuasion are required to establish the connection between politics and economics. In democracies it is 'obvious' that, to quote Harold Wilson, former Prime Minister of Great Britain, 'the standing of a Government and its ability to hold the confidence of the electorate of a General Election depends upon the success of its economic policy'. Given this dependence of political popularity on economic performance it is equally 'obvious' that a Government will try to manipulate its economic policy in such a way as to produce the most favourable outcomes just before election day, leaving the less favourable outcomes to occur at other times. Such behaviour on the part of governments is described in the literature as generating a 'political business cycle'.

But is what is obvious also true? Economists and political scientists have spent a great deal of time and effort correlating movements in political popularity – usually measured by the responses of voters to the question, 'If there was an election tomorrow which party/candidate would you vote for?' – with movements in key economic variables like the inflation rate, the unemployment rate and the rate of growth of real income (cf. Borooah and van der Ploeg (1984) for a survey of the literature). By and large the consensus has been that the proposition that political popularity depends on economic performance does have statistical support for a wide cross-section of countries across a variety of time periods. To this broad conclusion must, however, be added a number of caveats.

First, the relationship between political popularity and economic performance does not appear to be a stable one. Indeed there is evidence that the electorate's criteria for judging the economic performance of a government changes over time and, in particular, changes between periods relating to different governments. The economic problems that are highlighted – and by implication the problems that are underplayed – differ between governments. This may be partly due to differences in ideology and partly due to a desire to take advantage of existing economic conditions. To some extent this alters voters' perceptions of what is important; to another extent voters judge governments by the criteria governments themselves set. Both factors combine to produce changes over time in the criteria of economic success (cf. Borooah and van der Ploeg 1984; Butler and Stokes 1974).

Second, the idea that governments are rewarded for good times and punished for bad times assumes that voters are agreed on what constitutes good and bad times. In fact the most interesting issues in politics and economics are those on which voters, through differences of self interest among them, are divided (cf. Stigler 1973). Thus for example the proposal that tax relief on mortgage interest should be removed would be of no personal concern to persons living in publicly provided housing but would greatly alarm owner occupiers of dwellings.

Third, the existence of common interests among individuals leads to the formation of 'interest groups' which then attach themselves to the political parties they perceive as best representing their interests. The intensity of this attachment might increase in bad times (and decrease in good times) so that one consequence of economic hardship might not be a general reduction of support for the ruling party but an increase of support for each party in the class whose interests it represents and a decline in the support for each party in the class whose interests it does not represent (cf. Converse 1958).

Turning to the other side of the question, do governments in formulating their economic policies take into account the likely electoral impact of such policies and, in particular, do they use such policies to extract votes out of the electorate? The answer to this is not clear cut. At the macro level, in terms of generating the 'political business cycle', the evidence is mixed. The core proposition here is that *before* an election, a reflationary boost to the economy leads employment and incomes to rise and elicits a favourable response from voters. The price in terms of higher inflation is paid *after* the election, in response to which, deflationary measures (with a consequent fall in employment and incomes), to check the rise in prices, are taken. The process is repeated with the onset of the next election.

However, there is no systematic evidence for the existence of such a cycle. In the United States, for example, President Nixon, in 1972, attempted to generate an election year expansion, but there were no such attempts during President Ford's brief tenure in office. President Carter succeeded in running the cycle in reverse; by contrast President Reagan's policies were entirely consistent with the political business cycle model. The pursuit of politically motivated macro policies is made difficult by the economic constraints that governments face. This is especially true of small open economies. Thus in the United Kingdom, a major reflation would, with the shortest of delays, inevitably lead to a balance of payments crisis. A further difficulty is the length of voters' memories; if voters forget only slowly they may not be capable of being systematically deceived.

There is more evidence however, that, at the microeconomic level, governments further their political interests through the use of economic policies. Thus, within the context of a particular macroeconomic stance, a government may tailor the components of its micro-policy to favour certain sections of the electorate over others. For example in the public expenditure reduction programme of the 1979–83 Conservative government in the United Kingdom, expenditure on publicly provided housing was relatively hard hit. It is of course no surprise to learn that the users of such housing were not natural supporters of the Conservative party. Instances of such 'politically inspired' microeconomic policy can be multiplied without number. They have escaped the attention of economists, however, partly because of economists' preoccupation with macroeconomic, to the exclusion of microeconomic, policies and partly because economists have ignored the lack of homogeneity and the diversity of interests that exist in society. It is precisely upon the existence of such diversity that the logic of political intervention at the microeconomic level is based.

Indeed both Marx and Kalecki exposited theories of the political business cycle based upon the existence of societal conflict. Thus in the Marxist view 'the goal of macropolicy is not to eliminate the business cycle, but to guide it in the interests of the capitalist class' (Boddy and Crotty 1975, p. 10). Marx's own view was that the conflict between labour and capital over their shares of the national income would lead to cyclical booms and slumps. In the expansionary phase of the cycle the share of profits in national income would fall with a corresponding rise in the share of wages; to end this squeeze on profits capitalists would generate a slump by slowing down the rate of capital accumulation.

With the birth of Keynesianism grew the assumption that the maintenance of full employment was a proper and feasible objective of governments; this assumption if true would of course imply an end to business cycles. Kalecki however argued that

> the assumption that a government will maintain full employment if it only knows how to do so is fallacious . . . the class instinct of [business leaders] tells them that lasting full employment is unsound from their point of view and the unemployment is an integral part of the normal capitalist system. (Kalecki 1943, pp. 138–9 and pp. 140–41)

The concept and the role of the state was of crucial importance to Marx and Kalecki; both regarded the state as the institution, beyond all others, whose function it was to maintain and defend class domination and exploitation. However, although Kalecki's views were grounded in Marx's theory of conflict, Kalecki's reasons for the basis of the conflict were more social than economic. In his view, businessmen's dislike of lasting full employment was based on the fear that their social position would be undermined with the growing self-assurance of workers that such full employment would engender.

The views of Marx and Kalecki on business cycles are political in a broad sense in that the mechanism generating the cycles is class conflict between capitalists and workers, with government aiding the capitalist class and the working classes constraining their ability to do so. This insight regarding the importance of societal conflict in determining the course of the economy has unfortunately been ignored by mainstream economics which has instead sought to seek the interaction between politics and economics within the more narrow confines of electoral behaviour. However, there is hope for the future development of the subject in the form of three expanding areas of research.

First there is the growing interest, generated by Olson's (1965) seminal contribution, in the role of interest groups and the role of such groups in the shaping of economic policy. Borooah and van der Ploeg (1984), Hibbs (1982) and van Winden (1983) have analysed the differential impact of economic outcomes on different subgroups of voters and have pointed to the scope this offers for 'manipulative' economic policy, particularly at the microeconomic level. Second, there is a growing interest in the institutions that formulate and implement economic policy. A government's economic policy is not the output of a homogeneous entity but evolves through the resolution of conflict between its different branches. This could be between elected legislators and bureaucrats or between state and central governments or between a central policy-making department and other departments. In any event, to understand the making of economic policy one needs to understand the political culture within which it is embedded (cf. Heclo and Wildavsky 1981). Research in this area is still in its infancy and has focused mainly on the behaviour of the bureaucracy.

Finally, there is an attempt to examine the relation between economics and politics at the level of the international economy (cf. Frey 1984). This has looked at two areas: the role of international economic organizations, their organization and working and the political economy of tariffs and trade restrictions. From all this it would appear that the area of interaction between politics and economics is particularly worth cultivating. To make the harvest worthwhile however, it is important to avoid the sterility of identical, optimizing agents so beloved of neoclassical economics and take on board instead, the ideas of conflict and division within society, that an earlier generation of economists had provided.

See Also

- ▶ Collective Action
- ▶ Constitutional Economics

Bibliography

Boddy, R., and J. Crotty. 1975. Class conflict and macro-policy: The political business cycle. *Review of Radical Political Economics* 7(1): 1–19.

Borooah, V.K., and F. van der Ploeg. 1984. *Political aspects of the economy*, Department of Applied Economics occasional paper 55. Cambridge: Cambridge University.

Butler, D., and D. Stokes. 1974. *Political change in Britain*. London: Macmillan.

Converse, P.E. 1958. The shifting role of class in political attitudes and behaviour. In *Readings in social psychology*, ed. E.E. Maccoby, T.M. Newcomb, and E.L. Hartley. London: Methuen.

Frey, B.S. 1984. Public choice and global politics. *International Organization* 38: 199–223.

Heclo, H., and A. Wildavsky. 1981. *The private government of public money*. London: Macmillan.

Hibbs, D.A. 1982. Economic outcomes and political support for British Governments among occupational classes: A dynamic analysis. *American Political Science Review* 76: 259–279.

Kalecki, M. 1943. Political aspects of full employment. *Political Quarterly* 14: 322–331. Reprinted in Kalecki, M. *Essays on the dynamics of the capitalist economy, 1933–70*. Cambridge: Cambridge University Press.

Olsen, M. 1965. *The logic of collective action*. Cambridge: Cambridge University Press.

Stigler, G.J. 1973. General economic conditions and national elections. *American Economic Review* 63: 160–167.

van Winden, F.A.A.M. 1983. *On the interaction between state and private sector*. Amsterdam: North-Holland.

Pollution Haven Hypothesis

Arik Levinson

Abstract

The pollution haven hypothesis, or pollution haven effect, is the idea that polluting industries will relocate to jurisdictions with less stringent environmental regulations. Empirical studies of the phenomenon have been hampered by the difficulty of measuring regulatory stringency and by the fact that stringency and pollution are determined simultaneously. Early studies based on cross sections of data found no significant effect of regulations on industry locations. Newer studies that use panels of data to control for unobserved heterogeneity or instrumental variables to account for simultaneity have found statistically significant, reasonably sized effects.

Keywords

Environmental regulations; Fixed-effects models; Free trade; Heckscher–Ohlin trade theory; Inter-jurisdictional competition; Not in my backyard (NIMBY); Pollution abatement costs and expenditures; Pollution haven hypothesis; Porter hypothesis; Protectionism; Race to the bottom; Regulatory stringency; Tariffs; Unobserved heterogeneity

JEL Classifications

F1; F2; H73; Q5

The pollution haven hypothesis (or pollution haven effect) posits that jurisdictions with weak environmental regulations – ‘pollution havens’ – will attract polluting industries relocating from more stringent locales. The premise is intuitive: environmental regulations raise the cost of key inputs to goods with pollution-intensive production, and reduce jurisdictions’ comparative advantage in those goods. The Heckscher–Ohlin model provides the theoretical foundations by showing that regions will export goods that use locally abundant factors as inputs. Empirically, however, robust evidence that industries shift production to less stringent jurisdictions has proven elusive.

Econometric studies of the pollution haven effect have typically focused on reduced-form regressions of a measure of economic activity on some measure of regulatory stringency and other covariates:

$$Y_i = \alpha R_i + X_i' \beta_i + \varepsilon_i \tag{1}$$

where Y is economic activity, R is regulatory stringency, X is other characteristics that will affect Y, and ε is an error term. The pollution haven hypothesis is that estimates of $\partial Y/\partial R$ will be negative ($\widehat{\alpha} < 0$). The empirical literature contains a wide variety of implementations of (1). Some studies focus on international trade, where Y_i represents, say, net exports from country i, and the right-hand side contains country characteristics. Others focus on employment, foreign direct investment, or new manufacturing plant births. Equation (1) has also been used to examine the pollution haven hypothesis at the level of sub-national jurisdictions, such as US states or counties. Some studies have further disaggregated Y by industry, in the expectation that environmental regulations have a larger effect on polluting industries than on clean ones.

On the right-hand side of (1), finding an appropriate measure of regulatory stringency (R) is not simple. The problem is not merely one of collecting the appropriate data; merely conceiving of data that would represent R is difficult. What we want to know is how much more costly production is in a given jurisdiction relative to others, due to the jurisdiction’s environmental regulations. These environmental compliance costs could take many forms: environmental fees or taxes, permitting costs, regulatory delays, emissions limits that require installation of costly technology, the threat of lawsuits, product or process redesign, forgone output, and so forth. Some attempts to measure these costs involve creating indices by weighting various country or state characteristics such as environmental agencies’ budgets, public awareness of environmental problems, the number of international environmental agreements the country has joined, states’ congressional delegations’ voting on environmental issues, or other general indicators. Other studies have used measurements of pollution directly, arguing that, for example, high sulphur emissions are evidence of lax regulations. Studies based on US data have used measures of manufacturers’ pollution abatement expenditures by state or industry, using the US Census Bureau’s *Pollution Abatement Costs and Expenditures* (PACE) survey, which ran from 1973 to 1994 and resumed in 2005.

None of these measures of R is ideal for testing the pollution haven hypothesis. The compiled indices of stringency are inherently ad hoc, and typically not available in more than one cross section. Using pollution directly as a proxy for stringency is also problematic. High levels of pollution could be symptomatic of lax of regulations, or could mean that the jurisdiction, finding itself with a poor environment, must enact stringent regulations to reduce pollution. This is true in the United States, where counties that are out of compliance with national air-quality standards are required by the federal Clean Air Act to enforce stricter emissions laws. Even direct measures of abatement costs from the PACE are troublesome. States with the highest average abatement costs are those with the most polluting industrial compositions. Estimates of (1) in which average abatement costs proxy for R find that more polluting industries locate in places with higher abatement costs – the opposite of the pollution haven effect.

Even if we had available an ideal measure of regulatory stringency, R, two further econometric issues complicate estimates of Eq. (1): unobserved heterogeneity and simultaneity. The first problem is that some unobserved characteristics of the jurisdictions or industries being studied are likely to be correlated with both economic activity and regulatory stringency. A country with an unobserved comparative advantage in a polluting good (abundant high-sulphur coal or proximity to markets) is likely to both export that good and enact strict environmental regulations. This means that R and ε are correlated in (1), and estimates of $\hat{\alpha}$ will be biased. In fact, cross-section comparisons sometimes find that countries with higher stringency have more polluting activity, which is in turn easily mistaken for evidence of the Porter hypothesis that environmental regulations promote competitiveness (Porter and van der Linde 1995).

The simplest solution to the problem of unobserved heterogeneity is to estimate a panel-data version of (1) and include fixed effects by jurisdiction or industry, whatever the relevant unit of observation:

$$Y_{it} = \nu_i \ + \alpha R_{it} + X'\beta_{it} \ + \varepsilon_{it} \tag{2}$$

These fixed effects (ν_i) capture the unobserved characteristics of jurisdictions or industries that make them likely to have both strict environmental regulations and high levels of activity. However, including fixed effects requires panel data on regulatory stringency, which makes measuring stringency in the first place even more difficult.

The second econometric issue confronting estimates of (1) and (2) is that economic activity and pollution regulations may be determined simultaneously. The pollution haven hypothesis suggests that environmental regulations affect exports, but the reverse may also be true: exports may affect regulations. If trade increases incomes, and environmental quality is a normal good, trade could increase voters' demand for strict environmental regulations. Or, increased pollution caused by trade could increase local demand for strict environmental regulations. In theory the straightforward solution to this problem is to use instrumental variables. In practice this means finding instruments for a variable, R, that is difficult to measure in the first place. In the panel context (2), it means finding something that changes over time, is correlated with R_{it}, and is uncorrelated with ε_{it}.

The empirical studies that employ these techniques span more than 30 years, and are growing in number. While enumerating them here would be impractical, their broad lessons are becoming clear. The first generation of empirical work on the pollution haven hypothesis used cross sections of data and made no attempt to control for unobserved heterogeneity or simultaneity. Most of them found small insignificant effects of environmental regulations, a few found counter-intuitive positive effects, and none found robust significant support for the pollution haven hypothesis. This early literature is summarized in Jaffe et al. (1995, p. 157): 'Overall, there is relatively little evidence to support the hypothesis that environmental regulations have had a large adverse effect on competitiveness.'

In recent years, economists have begun to use panels of data and fixed-effects models to control for unobserved heterogeneity, and instrumental variables to control for simultaneity. In contrast to the earlier cross-section studies, this newer work has tended to find statistically significant, reasonably sized evidence of pollution havens. It is catalogued in detail by Brunnermeier and Levinson (2004), and summarized in Copeland and Taylor (2004, p. 48), who write that 'after controlling for other factors affecting trade and investment flows, more stringent environmental policy acts as a deterrent to dirty-good production'.

One example of this recent literature exploits the US Clean Air Act, which mandates that every county in the United States achieve the same minimum level of ambient air quality. Federal law requires counties that fail to attain this standard to implement more stringent regulations. A convenient aspect of this law for pollution haven research is that from the perspective of any single county the law is exogenous. Neither the law's first enactment in 1970 nor any subsequent tightening of the air quality standards has been a function of any one county's characteristics. This suggests that an indicator for whether a particular county is in compliance with the national standards makes a good instrument for the stringency of that county's environmental regulations. Non-compliance changes over time, is correlated (positively) with stricter regulations, and is unlikely to be correlated with ε_{it}. Using this strategy, Becker and Henderson (2000) find that a county's failure to meet the national air quality standards reduces the number of new plants being built by four heavily polluting industries by between 26 and 45 percent. Greenstone (2002) shows that these non-attainment counties had about 590,000 fewer jobs, $37 billion lower capital stock, and $75 billion lower output (in 1987 US dollars) between 1972 and 1987 than counties that met the national standards.

An important caveat should accompany findings of this type: they are positive, or descriptive, rather than normative. These tests of the pollution haven hypothesis merely measure whether industry relocates to less stringent jurisdictions; they have no welfare implications. Nevertheless, advocacy groups with widely varying agendas have seized on the issue. Some environmental groups express concern about pollution increases, resulting either from the trade-induced change in the pollution havens' industrial compositions or from the increase in overall economic activity due to trade. Manufacturing interests and labour unions in developed countries worry that the pollution haven effect means a loss of domestic profits and jobs. Free trade advocates fear that protectionist interests will use environmental regulations as a justification for trade barriers, or as a direct protectionist mechanism by lobbying for lower environmental standards as a form of subsidy to manufacturers. Anti-globalization protestors claim that trade liberalization will exacerbate all of these outcomes: degrading environmental quality in developing countries, weakening manufacturing in developed countries, and deterring all countries from setting sufficiently strict environmental standards.

In some cases these diverse parties have different or related interpretations of the pollution haven hypothesis. The most straightforward interpretation, represented by $\alpha < 0$ in Eqs. (1) and (2), is that environmental regulations cause polluting activity to shift to less stringent jurisdictions. Although virtually all of the empirical literature tests this descriptive hypothesis, much of the policy debate revolves around tangential issues with more normative implications.

One such related issue is whether trade liberalization exacerbates the pollution haven effect. Note the subtle difference. The straightforward pollution haven hypothesis is that environmental regulations affect trade. This extension claims that trade barriers disproportionately affect trade in polluting goods, and hence the environment. It seems that would be true only if the trade barriers had a larger effect on polluting industries than on clean industries. An empirical test of this extension would rewrite Eq. (2) to include trade barriers and an interaction between trade barriers and regulations:

$$Y_{it} = v_i + \alpha R_{it} + \gamma T_{it} + \theta R_{it} T_{it} + X' \beta_{it} + \varepsilon_{it} \qquad (3)$$

where T_{it} represents trade barriers such as tariffs (Ederington et al. 2004). The straightforward pollution haven effect is now $\partial Y = \partial R = \alpha + \theta T.$ The indirect effect of trade barriers on the pollution haven effect is $\partial\partial Y / [\partial T\, \partial R] = \theta.$

Given the difficulties in measuring both regulatory stringency and trade barriers, and the likely endogeneity of both, few studies have attempted to estimate this indirect effect of trade liberalization on pollution havens. Nevertheless, it is important to be clear that the basic empirical estimates of the pollution haven effect do not address this more complex extension.

A second concern related indirectly to the pollution haven hypothesis is that governments will engage in inefficient competition to attract polluting industries by weakening their environmental standards. A welfare-maximizing government should set standards so that the benefits justify the costs at the margin. This does not mean that environmental standards will be equal everywhere. Jurisdictions have different assimilative capacities, costs of abatement, and values regarding the environment. So heterogeneity in pollution standards is to be expected, and by extension industry migration to less stringent jurisdictions does not necessarily raise efficiency concerns.

There might be cause for concern, however, if jurisdictions compete for investment from polluting industries by setting environmental regulations below Pareto-efficient levels. They might do so, for example, if there were cross-border spillovers, and the benefits of hosting a polluting manufacturer outweighed the *local* costs. Alternatively, if the industry is concentrated and pays rents to outside shareholders, jurisdictions may compete away their ability to capture some of the industry's rents. In these types of case, countries may lower their regulations below the Pareto-optimal levels in a 'race to the bottom' in environmental standards. Depending on the costs and benefits of hosting a polluting industry, they may also raise their standards above the Pareto-optimal levels in what has been called the 'not-in-my-backyard' (NIMBY) phenomenon. Levinson (2003) summarizes the theoretical and empirical literature on inter-jurisdictional environmental competition.

These questions of trade liberalization and inter-jurisdictional competition, however, extend the central issue of the pollution haven hypothesis. Most empirical studies of the pollution haven hypothesis ask the straightforward, descriptive question: have pollution-intensive industries become concentrated in jurisdictions with less stringent regulations? Early analyses based on cross sections of data typically found that environmental regulations had small or statistically insignificant effects on industry location. However, recent studies using panel data to control for unobserved heterogeneity or instrumental variables to control for the simultaneity of regulations have found statistically significant, reasonably sized pollution haven effects.

See Also

▶ Environmental Economics
▶ International Trade Theory
▶ Strategic Trade Policy
▶ Trade Policy, Political Economy of

Bibliography

Becker, R., and J. Henderson. 2000. Effects of air quality regulations on polluting industries. *Journal of Political Economy* 108: 379–421.

Brunnermeier, S., and A. Levinson. 2004. Examining the evidence on environmental regulations and industry location. *Journal of Environment & Development* 13: 6–41.

Copeland, B., and M. Taylor. 2004. Trade, growth, and the environment. *Journal of Economic Literature* 42: 7–71.

Ederington, J., Levinson, A., and Minier, J. 2004. Trade liberalization and pollution havens. *Advances in Economic Analysis & Policy* 4(2) Article 6. Berkeley Electronic Press.

Greenstone, M. 2002. The impacts of environmental regulations on industrial activity: Evidence from the 1970 and 1977 Clean Air Act amendments and Census of Manufactures. *Journal of Political Economy* 110: 1175–1219.

Jaffe, A., S. Peterson, P. Portney, and R. Stavins. 1995. Environmental regulations and the competitiveness of US manufacturing: What does the evidence tell us? *Journal of Economic Literature* 33: 132–163.

Levinson, A. 2003. Environmental regulatory competition: A status report and some new evidence. *National Tax Journal* 56: 91–106.

Porter, M., and C. van der Linde. 1995. Toward a new conception of the environment–competitiveness relationship. *Journal of Economic Perspectives* 9(4): 97–118.

U.S. Census Bureau. Various years. *Pollution abatement costs and expenditures, MA200*. Washington, DC: U.S. Government Printing Office.

Pollution Permits

Ted Gayer

Abstract

Government can reduce pollution by issuing permits to polluters in numbers below existing emission levels. Under a tradable permit programme, a firm with high abatement costs can buy permits from another firm with low abatement costs, leading to a reduction in the total cost of abating relative to a system where reduction levels are strictly assigned. For tradable permits to work effectively, the emissions must come from discrete point sources and be relatively easy to monitor. Aside from issuing the permits, the government's role is to enforce compliance and establish optimal penalties for non-compliance.

Keywords

Auction hot spot; Carbon emissions; Clean Air Act Amendments of 1990.; Coase Theorem; Externalities; Market failure; Pollution permits; Property rights; Transaction costs

JEL Classifications

Q5

The government issues pollution permits to designate how many units of a given pollutant the permit owner is legally allowed to emit in a given period. The government can therefore reduce pollution from these sources by setting the total number of permits below their total existing emission levels. The cost savings of this approach result from allowing the pollution permits to be traded (Dales 1968). Under such a tradable permit programme (also known as a cap-and-trade programme), a firm that has high abatement costs can buy permits from another firm that has low abatement costs, leading to a reduction in the total cost of abating relative to a system where reduction levels are strictly assigned. For tradable permits to work effectively, the emissions must come from discrete point sources and be relatively easy to monitor. Aside from issuing the permits, the government's role is to enforce compliance and establish optimal penalties for non-compliance.

Tradable pollution permits can help address welfare losses caused by pollution. In a free market system goods are exchanged voluntarily. Buyers and sellers engage in trade only if both parties believe they will benefit from the exchange. These trades are coordinated by market prices, which convey information to all parties on the demand for the good and the cost of supplying the good. This system of mutual improvement results in an efficient allocation. However, inefficiency may result if a voluntary transaction between two parties imposes involuntary costs on a third party. These third-party costs are known as externalities.

The root of this market failure is that there are no clear property rights for the surrounding air. Consider an example of a firm which emits air pollution that imposes costs on its neighbours. If the firm's neighbours owned the rights to clean air, then the firm would need to compensate the neighbours in order to use the air in its production process. Similarly, if the firm owned the rights to pollute the air, the neighbours could pay the firm to reduce its emissions. In either setting, the market would incorporate both the costs of pollution to the neighbours and the benefits of pollution to the firm, resulting in an efficient outcome. Indeed, the key interpretation of the Coase Theorem (Coase 1960) is that efficiency results no matter who is legally assigned the property right to the air, so long as free exchange is possible and there are no transaction costs.

The necessary condition of no (or even low) transaction costs is likely to be violated when there are many sources of pollution and when many people bear the external costs of the pollution. This presents an economic justification for government involvement, since the absence of a working market for clean air will lead to an externality-induced inefficiency.

If high transaction costs preclude efficiency-enhancing bargaining between third parties and polluters, then the government can assume the role of the property right owner for the air. Because the 'government' is not an individual cost-bearing entity in the same manner as the affected third parties, and because government agents may have goals other than efficiency, it is not assured that government regulation will lead to an efficient outcome. Ideally, the role of government would be to assess the external costs associated with the production process and to determine the pollution reduction level that maximizes net benefits. The government could then issue tradable permits that yield this efficient level of pollution reduction.

It is undoubtedly difficult to determine the efficient amount of pollution reduction. However, no matter which target level is chosen, a system of tradable permits can help achieve the goal in the least costly way. In a cap-and-trade system, a firm with a high cost of reducing an additional unit of emissions could purchase a pollution permit from a firm with a lower marginal abatement cost. This trading will continue until the marginal abatement costs are equal across firms, thus minimizing the total cost. The cost-savings occur no matter if the government initially gives out the permits to firms for free (known as grandfathering), or if the government decides to auction the permits. Given a competitive market for permits, the initial allocation of permits has only distributional consequences, not efficiency consequences.

The freedom to trade permits across firms or to bank (or even borrow) permits across time results in cost savings without violating the long-term total pollution reduction goal. Additionally, by creating a property right for pollution, a cap-and-trade system establishes a market price for pollution and therefore provides firms with an incentive to find less expensive ways to reduce emissions (Carlson et al. 2000). In contrast, a regulation that rigidly sets technological standards for a firm does not provide this incentive.

Some environmental problems are difficult to address with tradable permits. For example, if the marginal damage of emissions varies by location (for example, due to variation in existing ambient concentrations or due to differences in the number of people exposed to the pollutant), then a tradable permit system might shift emissions from a low-damage to a high-damage location and thus increase total damages. A congestion of emissions in one location, known as a 'hot spot', could result in greater damage than if pollution were reduced uniformly across polluters. A cap-and-trade system can address this problem by making the required number of permits per unit of emissions a function of the marginal damage, or by establishing separate permit markets by region. However, these options do add a level of complexity. In addition, in order for a tradable permit market to work efficiently, it must be a competitive market composed of informed buyers and sellers. While tradable permits can minimize costs, in practice such programmes are grafted on to existing command-and-control regulations, which can affect the cost savings (Hahn 1989).

Since around 1985 the United States has adopted a number of tradable permit programmes to address a variety of pollution problems (Stavins 2000). These include the phase-out of leaded gasoline, chlorofluorocarbon trading, the Regional Clean Air Incentives Market (RECLAIM) to address sulfur dioxide and nitrogen oxides, and the recent Nitrogen Oxides State Implementation Plan (SIP) Call. The most notable example of a cap-and-trade system is the sulphur dioxide programme for electricity generating units, which was enacted under the Clean Air Act Amendments of 1990. This programme has achieved its pollution-reduction goals with estimated cost savings of approximately $1 billion a year compared with costs under a hypothetical command-and-control regulatory alternative.

See Also

- ► Auctions (Empirics)
- ► Environmental Economics
- ► Environmental Kuznets Curve

Bibliography

Carlson, C., D. Burtraw, M. Cropper, and K. Palmer. 2000. Sulfur dioxide control by electric utilities: What are the gains from trade? *Journal of Political Economy* 108: 1292–1326.

Coase, R. 1960. The problem of social cost. *Journal of Law and Economics* 3(1): 1–44.

Dales, J. 1968. *Pollution, property and prices*. Toronto: Toronto University Press.

Hahn, R. 1989. Economic prescriptions for environmental problems: How the patient followed the doctor's orders. *Journal of Economic Perspectives* 3(2): 95–114.

Stavins, R. 2000. Market-based environmental policies. In *Public policies for environmental protection*, 2nd ed., ed. P. Portney and R. Stavins. Washington, DC: Resources for the Future.

Pontryagin's Principle of Optimality

M. I. Zelikin

JEL Classifications
C6

The meaning of the word 'economics' is closely related with that of 'optimality'. It is for this reason that methods used in the theory of optimal control find their natural practical application in economics.

In this entry we deal with the statement of Pontryagin's maximum principle and give an exposition of the results and the perspectives of its applications to macroeconomic optimizational problems. We are concerned with two lines in the development of macroeconomics – that of Ramsey and that of von Neumann. Pontryagin's maximum principle embraced both lines – they now coexist in the principle, being inseparable and yet unmergeable. We begin with the classical formulation of Pontryagin's principle.

Let the state of the given system be described by the vector $x = (x_1, \ldots\ldots, x_x)$; $x \in X \subset \mathbb{R}^x$ (X is an open domain). Control is described by the vector $u = (u_1, \ldots., u_y)$; $u \in U \subset \mathbb{R}^y$. The independent variable t is time. For control one can choose any piecewise continuous function $u(t)$, whose values belong to U. The dynamics of the system are described by the equations $\dot{x}_i = \varphi_i(t, x, u(t)), (i - 1, \ldots, n); x(t_0) = a$. The pair consisting of the control $u(t)$ and the corresponding path $x(t)$ is called the process. A smooth manifold M in the space (t, x) is given, and the first hitting time of this manifold M is taken as the moment of termination of the process. The hitting time is the moment of first arrival of the point x at the manifold M, i.e. $T = \inf\{t| x(t) \in \mathrm{M}\}$. In the case when M is the hyperplane $t = T =$ Const one says that this is a fixed time and free end problem. The criterion is the functional

$$x_0 = F(T, x(T)) + \int_{t_0}^{T} f(t, x(t), u(t))dt \to \sup.$$

In the case $f \equiv 0$ and $T =$ Const the functional x_0 is said to be terminal. It is assumed that the functions f, F and φ are smooth.

To formulate Pontryagin's maximum principle let us consider a dual (or an adjoint) vector $\psi = (\psi_{0,}\psi_1, \ldots. , \psi_n)$ and the Pontryagin function

$$H(t, \psi, x, u) = \psi_0 f(t, x, u) + \sum_{\alpha=1}^{n} \psi_\alpha \varphi_\alpha(t, x, u).$$

Pontryagin's Maximum Principle

If $u^*(t)$, $x^*(t)$ is the optimal process, then there exists a nontrivial, continuous vector- function $\psi(t) = (\psi_{0,}\psi_1(t), \ldots. , \psi_n(t))$ with the following properties.

1.The adjoint equations:

$$\dot{\psi}_0 = 0;$$
$$\dot{\psi}_i = -\frac{\partial H}{\partial x_i}(t, \psi, x^*(t), u^*(t)), \quad (i = 1, \ldots, n).$$

2.The transversality conditions:

$\psi_0 \geq 0$ the $(n + 1)$-dimensional vector

$$\left\{\psi_1(T) - \psi_0 \frac{\partial F}{\partial x_1}(T, x^*(T)), \ldots., \psi_n(T) = \psi_0 \frac{\partial F}{\partial x_n}(T, x^*(T)), \times - H(T, \psi(T), x^*(T), u^*(T)) - \psi_0 \frac{\partial F}{\partial T}(T, x^*(T))\right\}$$

is orthogonal to the manifold M at the point $(T,x^*(T))$

3.The maximum condition:

$$\max_{u \in U} H(t, \psi(t), x^*(t), u) \\ = H(t, \psi(t), x^*(t), u^*(t)).$$

In the case of fixed time and free end, the transversality conditions reduce to

$$\psi_0 \geq 0, \quad \psi_i(T) \\ = \psi_0 \frac{\partial F}{\partial x_i}(t, x^*(T)), \quad (i = 1,\ldots,n)$$

(if $F = 0$, we have $\psi_i(T) = 0, (i = 1, \ldots, n)$). Let us remark that the vector ψ is defined up to multiplication by a positive constant, and in the case $\psi_0 \neq 0$ it can be normalized by dividing by ψ_0. As soon as the optimal value u^* at the point (t, x) depends only on that point, we can seek the optimal control as a feedback control, i.e. in the form $u^* = u^*(t, x)$. This function defines the optimal control at each point of the space (t, x), and thus it is called the optimal synthesis. The variables H and ψ can also regarded as functions of t and x. Let us denote the optimal value of the functional, corresponding to the initial point (t, x), by $x_0(t, x)$. This function is called Bellman's function.

The main idea of the economic interpretation of Pontryagin's maximum principle (which goes back to L.V. Kantorovich) is to consider the variables ψ_i as shadow prices. To explain, let us assume that the problem is regular ($\psi_0 \neq 0$), and that the optimal synthesis $u^*(t, x)$ and the dual variables $\psi(t, x)$ are smooth. In this case $x_0(t, x)$ is also smooth and Bellman's equation

$$\max_{u \in U} \left\{ \frac{\partial x_0}{\partial t}(t,x) + f(t,x,u) + \sum_{\alpha=1}^{n} \frac{\partial x_0}{\partial x_\alpha}(t,x)\varphi_\alpha(t,x) \right\} = 0$$

is fulfilled. The relationship between Bellman's equation and Pontryagin's maximum principle can be expressed by the equations

$$\frac{\partial x_0}{\partial x_i}(t,x) = \frac{\psi_i(t,x)}{\psi_0(t,x)}; \quad \frac{\partial x_0}{\partial t}(t,x) = -\frac{H(t,x)}{\psi_0(t,x)},$$

i.e. the normalized value of the dual variable ψ_i/ψ_0 is the marginal effect of the factor x_i on the optimal value of the functional x_0 and that is exactly the shadow price of x_i. The economic meaning of Pontryagin's maximum principle is as follows. For the optimal process $u^*(t)$, $x^*(t)$ there exist shadow prices ψ the adjoint equations and the transversality conditions being fulfilled, such that the optimal value of the control $u^*(t)$ at each moment t maximizes the flow of the profit, which is calculated in accordance with the shadow price. It is worth remembering, that in the irregular case ($\psi_0 = 0$), as well as in the case of discontinuous optimal synthesis $u^*(t,x)$ Bellman's function $x_0(t, x)$ is often nonsmooth, in spite of all the functions defining the statement of the problem being smooth. Bellman's equation breaks down, but Pontryagin's maximum principle is fulfilled. In that case the notion of 'prices' loses its natural meaning. The search for general enough conditions, guaranteeing the smoothness of Bellman's function, is a difficult and only partially explored mathematical problem.

The creation of Pontryagin's principle stimulated the two aforementioned lines of macroeconomics. Before listing the corresponding results, let us note some significant obstacles in the way of application of these optimizing methods to mathematical economics. To formulate an optimization problem, we have to choose a criterion. It is only natural to take as a criterion some function of the final state $x(T)$ or the profits over some interval of the time $[t_0, T]$. But the choice of the moment T (the horizon of the plan) is arbitrary from an economic point of view. Meanwhile it is highly desirable to define economically reasonable behaviour independently of such arbitrariness. Two approaches to overcome this obstacle are known–that of F.P. Ramsey (and his collaborator J.M. Keynes) and that of J. von Neumann.

Ramsey's approach is to take eternity as the horizon of the plan. He applied the calculus of variations, which can be regarded as a version of Pontryagin's maximum principle, to the problem of resource allocation between consumption and saving, aiming to maximize the benefits of society during the entire infinite period of its possible existence, and proved the Golden Rule of saving. From the mathematical point of view the problem is to minimize the integral over the half-open interval $[t_{0,\infty}]$ from the difference between absolute welfare (Bliss) and immediate welfare (the utility function), which tends to Bliss and depends on the solution of the differential equation containing the policy of saving as control. The principal part of the right-hand side of this equation is the production function. The naivety of this model lies in the conception of stationary and absolutely stable economic Universe, rather than in the assumption of the possibility of complete aggregation. I hope to be indulged in using such unusual (in economic context) terms as 'Universe'. But in fact we deal with the closed macroeconomic models, purporting to describe all basic economic phenomena, and in this sense the situation is closely related to that of physical models of the Universe; hence the reason for the proposed usage.

Later on there were attempts to modernize this model and to make it nonstationary. On the one hand, Hicks, Harrod and Solow among others varied the production function, aiming to include in it the effect of technical change. On the other hand, T.C. Koopmans broke off the relationship between the production and the utility functions (which was so essential in the views of Ramsey) and introduced a discounting factor in the integrand which guaranteed the convergence of the functional for any choice of control. The stationarity of Ramsey's economic Universe was slightly shaken.

The path of the Golden Rule in such models appears to be a singular path of Pontryagin's maximum principle. The path of the principle is called singular, if the maximum of the Pontryagin's function (3°) at all points of this path is attained at several distinct points of the set U. In the case of problems which are linear in the control the nonsingular (band-bang) control uses only the extreme points of the set U. Such a control corresponds to the economic policy with sharp changes (switches). The characteristic feature of bang-bang optimal control is instantaneous switches from one vertex of the polyhedron U onto another. On the contrary, the singular control, using the internal points of U, as a rule does not need switches and in this sense seems to be more acceptable than the bang-bang control from the economic point of view.

The application of Pontryagin's principle to such models calls for its generalization to the case of infinite time intervals. More precisely, the transversality conditions at infinity need generalization, and only those. But even that question turns out to be difficult enough. There were incorrect attempts to formulate these conditions in the form: $\psi(t) \to 0$ for $t \to \infty$. For the case of linear differential constraints Aubin and Clarke found out and proved a correct version of transversality conditions, which requires the convergence of the integral of $|\psi(t)|^q e^{\partial t}$ for some q and δ. A complete and correct solution of the problem of transversality conditions at infinity for nonlinear differential constraints is still absent.

Another line of evolution of macroeconomics begins with the work of von Neumann. He introduced a discrete model of the expansion of production, which was defined by two given matrices – that of input and that of output. Von Neumann seeks the balanced development of economics when the input vector is proportional to the output one. Among these rays he seeks those yielding maximal growth. He introduces dual variables – prices of the optimal plan – and gave optimality conditions in dual terms.

Later on it turned out (in accordance with the hypothesis of Samuelson) that the ray of maximal growth (now called) 'Neumann's ray' or 'the turnpike') plays the leading part in exploring the optimal paths of this model. The corresponding theorems (turnpike theorems) assert that this ray defines the asymptotic behaviour of the optimal paths, in the increase of the horizon of the plan (for $T \to \infty$), independently of the choice of terminal functional, and it is precisely this fact which makes it possible to overcome the aforementioned obstacle. Continuous versions of the turnpike theorems, which require Pontryagin's principle or the related methods for their proofs, were obtained by, among others, A.N. Ducalov, A.E. Ilutovich and L.F. Zelikina. Let us note that the turnpike in these models is the singular path of the principle. In the work of Zelikina, for certain optimal resource allocation problems, Neumann's concept of optimal policy, independent of the choice of functional, was brought to its logical conclusion. By constructing the optimal synthesis in n-dimensional space, it was shown that the shadow prices for increasing (with $T \to \infty$) initial segments of optimal paths are invariant relative to such a choice. The search for a complete system of invariants of optimal synthesis relative to a choice of functional (in some appropriate class of the latter) for the general economical optimizational problem still remains an open question.

A hint for an infinite-dimensional version of the theory of duality is contained in the formulation of Pontryagin's maximum principle itself. A.M. Ter-Krikorov (1977) introduces the dual problem to the linear optimizational problem, in which the dual variables ψ_i turn into the state variables and vice versa.

The techniques of turnpike theorems and Pontryagin's principle gives rise to a series of optimization models, which become more and more universal and finally develop into the concept of expanding economic Universe. The economic analogue of the physical concept of the oscillating Universe is yet developed only on the phenomenological level, in spite of the empirical evidence for the corresponding economic phenomenon. The reason for this fact is (as it seems) the lack of satisfactory optimization models taking into account the specific effect of money.

It is worth noting that, like the physical models of Universe, all these concepts of economic Universe are devoid of its substance – thought and ethics – which naturally bring the question out of the competence of pure economics. But without this substance, the economical Universe, as soon as it claims to be universal, cannot be explained in principle.

See Also

- ▶ Hamiltonians
- ▶ Stochastic Optimal Control

Bibliography

Aubin, J.P., and F.H. Clarke. 1979. Shadow prices and duality for a class of optimal control problems. *SIAM Journal on Control and Optimization* 17 (5): 567–586.

Diukalov, A.N. and Iliutovich, A.E. 1973. Asimptoticheskie svoistva optimal'nikh traektorii ekonomicheskoi dinamiki (Asymptotic properties of optimal trajectories in economic dynamic). *Automatika i Telemakhanika* No. 3.

Harrod, R.F. 1937. Remark. *American Economic Review* 27.

Hicks, J.R. 1932. *The theory of wages*. London: Macmillan.

Koopmans, T.C. 1957. *Three essays on the state of economic science*. New York: McGraw-Hill.

Ramsey, F.P. 1928. A mathematical theory of saving. *Economic Journal* 38: 543–559.

Solow, R.M. 1957. Technical change and the aggregate production function. *The Review of Economics and Statistics* 39: 312–320.

Ter-Krikorov, A.M. 1977. *Optimal'noe Upravlenie i Matematicheskaia Ekonomika*. Moscow: Nauka.

von Neumann, J. 1945–1946. A model of general economic equilibrium. *Review of Economic Studies* 13(1): 1–9.

Zelikina, L.F. 1975. Universal manifolds and turnpike theorems for a class of optimal control problems. *Soviet Matematics Doklady* 16 (5): 1136–1140.

Zelikina, L.F. 1977. Mnogomernyi sintez i teoremy o magistrali v zadachkh optimal'nogo upravleniia (Multi-dimensional feed-back control and the turnpike theorems in the optimal control problems). In *Veroiatnostnye Problemy Upravleniia v Ekonomike* (Stochastic problems of control in economics). Moscow: Nauka.

Poor Law, New

George R. Boyer

Abstract

The New Poor Law refers to the system of local public assistance in England and Wales initiated by the passage of the 1834 Poor Law Amendment Act. This act attempted to restrict relief outside of workhouses for the able-bodied, but was evaded for three decades. The Crusade against Outrelief of the 1870s marked a major shift in administration and the increased use of workhouse relief. Numbers on relief fell sharply thereafter, although the elderly continued to rely heavily on the Poor Law. The Liberal welfare reforms of 1906–11 paved the way for the 1948 abolition of the Poor Law.

Keywords

Charity Organization Society; New Poor Law; Outdoor relief; Poor rate

JEL Classifications

N4

The New Poor Law refers to the welfare policy in England and Wales initiated by the passage of the Poor Law Amendment Act in 1834. All destitute individuals were eligible for poor relief from their local Poor Law union. Those granted assistance were either given cash or in-kind payments in their homes (outdoor relief) or were relieved in workhouses (indoor relief). Although the Poor Law remained in existence until 1948, the Crusade against Outrelief in the 1870s and the adoption of the Liberal welfare reforms in the decade before the First World War significantly reduced its role as a safety net for the poor.

The Poor Law Amendment Act was an outgrowth of the Report of the Royal Commission to Investigate the Poor Laws (1834), which called for sweeping reforms to the existing system of poor relief, including the grouping of parishes into Poor Law unions, the abolition of relief for the able-bodied and their families outside workhouses, and the appointment of a centralized Poor Law Commission to direct the administration of relief. The Act implemented some of the report's recommendations, but left the regulation of outdoor relief to the Poor Law Commissioners.

By 1839 most rural parishes had been grouped into Poor Law unions, which had built or were building workhouses. However, the Poor Law Commission met with strong opposition when it attempted to set up unions in the industrial north, and the implementation of the New Poor Law was delayed in several industrial cities. The Commission and its 1847 replacement, the Poor Law Board, issued orders in 1842, 1844 and 1852 to restrict the payment of outdoor relief to able-bodied males, but these were evaded by both rural and urban unions. Thus, while real per capita relief expenditure fell by 43 per cent from 1831 to 1841, and remained at least 20 per cent below its 1831 level for the remainder of the 19th century (see Table 1), many Poor Law unions continued to grant outdoor relief to needy able-bodied males after 1834 (Rose 1970; Digby 1978). Data for three London parishes and six provincial towns in the years around 1850 indicate that large numbers of prime-age males continued to apply for relief, and that a majority of those assisted were granted outdoor relief (Lees 1998). The Poor Law played an important role in assisting the unemployed and their families in urban districts during cyclical downturns (Boot 1990; Boyer 2004). Moreover, the New Poor Law, like its predecessor, provided a major source of support for the non-able-bodied poor. From the 1840s to the 1860s, in much of rural England a large share of those aged 70 and over received regular poor relief payments, although these often did not provide full maintenance (Thomson 1984).

Data on the number of persons receiving poor relief are available for two days a year, 1 January and 1 July, beginning in 1849; the official estimates of the annual number relieved in Table 1 are the average of the number relieved on these two dates. Studies conducted by Poor Law administrators in 1892 and 1906–07 found that the day counts significantly underestimated the number assisted during the year. The 'revised' estimates in Table 1 are based on these studies, and assume that the ratio of actual to counted paupers was 2.24 for 1851–96 and 2.15 for 1901–11. These estimates indicate that from 1850 to 1870 about ten per cent of the population was assisted by the Poor Law each year. Lees (1998) contends that over a three-year period as much as 25 per cent of the population made use of the Poor Law.

Relief expenditures were financed by a local property tax, known as the poor rate. Up to 1865, each parish within a Poor Law union was responsible for relieving its own poor. As a result, tax rates were often significantly different across parishes within Poor Law unions, and were especially high in working-class districts. Economic crises put enormous financial strain on parishes that were already poor. The 'basic weaknesses' of the poor relief system were exposed in the 1860s, when the Poor Law 'was subjected to an almost continual series of shocks' (Rose 1981). The two major shocks of the decade were the Lancashire

Poor Law, New, Table 1 Relief expenditures and numbers on relief, 1831–1936

Year	Expenditures on Relief (1,000 £s)	Real expenditure per capita 1831 = 100	Number relieved (official) 1,000s	Share of population relieved (official)	Number relieved (revised) 1,000s	Share of population relieved (revised)	Share of paupers relieved indoors
1831	6,799	100.0					
1836	4,718	75.1					
1841	4,761	57.2					
1846	4,954	64.3					
1851	4,963	62.8	941	5.3	2,108	11.9	12.1
1856	6,004	57.5	917	4.9	2,054	10.9	13.6
1861	5,779	55.6	884	4.4	1,980	9.9	13.2
1866	6,440	60.2	916	4.3	2,052	9.7	13.7
1871	7,887	67.9	1,037	4.6	2,323	10.3	14.2
1876	7,336	58.2	749	3.1	1,678	7.0	18.1
1881	8,102	64.0	791	3.1	1,772	6.9	22.3
1886	8,296	66.7	781	2.9	1,749	6.4	23.2
1891	8,643	66.9	760	2.6	1,702	5.9	24.0
1896	10,216	78.4	816	2.7	1,828	6.0	25.9
1901	11,549	78.5	777	2.4	1,671	5.2	29.2
1906	14,036	89.8	892	2.6	1,918	5.6	31.1
1911	15,023	86.7	886	2.5	1,905	5.3	35.1
1921	31,925	69.7	627	1.7			35.7
1926	40,083	118.9	1,331	3.4			17.7
1931	38,561	124.0	1,090	2.7			21.5
1936	44,379	153.5	1,472	3.6			12.6

Notes: Relief expenditure data are for the year ended on 25 March. In calculating real per capita expenditures, I used cost of living and population data for the previous year

Sources: Columns 1, 3, 4, and 7 from Williams (1981). Estimates in columns 5 and 6 constructed by the author following Lees (1998). Estimates in column 2 constructed by the author

cotton famine of 1862–4 and the East London crises of 1860–1 and 1867–9. The collapse of raw cotton imports from the United States during the American Civil War forced Lancashire cotton textile factories to shut down or severely curtail production. The resulting unemployment caused a huge increase in demand for relief, which the hardest-hit parishes were unable to meet, and led several Poor Law unions to appeal to private relief committees for charitable assistance. During the severe winters of 1860–1, 1867–8 and 1868–9, Poor Law unions in London's East End were also forced to turn to private charities for assistance in meeting the high demand for relief.

The problems associated with Poor Law finance led parliament to adopt the Union Chargeability Act in 1865, and similar acts relating to London in 1867, 1869 and 1870. These acts placed the cost of poor relief on the Poor Law union rather than on each parish within it, and thus shifted a large share of the cost of relief from working-class parishes (which had low tax bases and many paupers) to middle-class parishes (with higher tax bases and fewer paupers). The tax-shifting eased the financial burdens that had plagued the Poor Law, but also led to the revolt of middle- class taxpayers in many areas.

The Union Chargeability Act was one of the catalysts of the Crusade against Outrelief in the early 1870s. Encouraged by the Local Government Board (LGB), Poor Law unions throughout England and Wales curtailed outdoor relief for all types of paupers. In December 1871 the LGB issued a circular concluding that generous outdoor relief was destroying self-reliance among the poor. In the circular's words: ‘a certainty of obtaining outdoor relief in his own home whenever he may ask for it extinguishes in the mind of the labourer all motive for husbanding his resources, and induces him to rely exclusively upon the rates instead of upon his own savings for such relief as he may require' (quoted in Englander 1998, p. 107). The Charity Organization Society (COS), founded in 1869, aided the Board in convincing the public of the need for reform. It argued that most low-skilled workers earned enough to be able to set aside some income in anticipation of future interruptions in earnings caused by unemployment or sickness. The LGB and the COS maintained that the restriction of outdoor relief would improve the moral and economic condition of the poor in the long run. The COS also believed that most applicants for relief would refuse to enter workhouses and would remove themselves from relief roles, so that a shift from outdoor to workhouse relief would significantly reduce Poor Law expenditures. Most Poor Law unions found it difficult to resist a policy that promised to raise the morals of the poor *and* reduce taxes (MacKinnon 1987).

The effect of the Crusade against Outrelief can be seen in Table 1. Real per capita relief expenditures and the share of the population receiving relief both fell sharply from 1871 to 1876. The decline in numbers on relief was largely a result of the deterrent effect of the workhouse: as the COS predicted, many of those offered indoor relief refused it. From 1871 to 1881, the number of paupers receiving outdoor relief fell by 282,000 (a 33 per cent decline), while the number relieved in workhouses rose by only 21,000.

Real per capita relief expenditures increased after 1876, mainly because the Poor Law provided increasing amounts of medical care for the poor. Otherwise, the role played by the Poor Law declined in the last quarter of the 19th century. The share of the population receiving relief fell from seven per cent in 1876 (revised estimates) to 5.2 per cent in 1901. The decline was due in large part to improvements in living standards, which increased workers' ability to save and to join friendly societies – mutual help associations providing sickness, accident, death, and (sometimes) old age benefits. However, part of the decline in numbers on relief was a result of the Crusade against Outrelief and of a change in the attitude of the poor towards relief. Prior to 1870, a large share of the working class regarded access to public relief as an entitlement, although they rejected the workhouse as a form of relief. Partly as a result of COS propaganda, by the end of the century most within the working class viewed poor relief as stigmatizing, and went to great lengths to avoid applying for relief. Thus, the decline in the share of the population receiving poor relief from 1871 to 1901 overestimates the decline in the share living in poverty.

One section of the working class continued to rely heavily on the Poor Law – the elderly. Table 2 shows that, for the 12-month period from March 1891 to March 1892, 29.3 per cent of those aged 65 and over received poor relief, as compared with 5.1 per cent of children and 3.7 per cent of those aged 16–64. Most elderly paupers received only partial maintenance, which they combined with wage income, savings, friendly society or trade union benefits, and help from relatives or friends to achieve a subsistence income. The ability of the elderly to support themselves declined with age; Booth (1894) estimated that 40 per cent

Poor Law, New, Table 2 Pauperism in the early 1890s: 1 January 1892 and March 1891–March 1892

		1 January			12 months' count		
Ages	Population 1891	Indoor paupers %	Outdoor paupers %	Total paupers %	Indoor paupers %	Outdoor paupers %	Total paupers %
Under 16	10,762,808	0.5	1.6	2.1	1.0	4.1	5.1
16–64	16,867,116	0.5	0.7	1.2	1.4	2.3	3.7
65 and older	1,372,601	4.6	14.9	19.5	8.3	21.0	29.3
Total	29,002,525	0.7	1.7	2.4	1.3	3.8	5.4

Source: Report of the Royal Commission on the Aged Poor, Parliamentary Papers (1895), vol. XIV, pp. xii–xiii of those aged 75 and over received poor relief, as compared to 20 per cent of those aged 65–70.

Despite improvements in living standards, many manual workers still experienced 'acute financial' distress at some point in their lives (Johnson 1985). The inability of low-skilled workers to protect themselves from financial insecurity was the catalyst for the Liberal welfare reforms, several pieces of social welfare legislation adopted between 1906 and 1911. Acts of 1906 and 1907 provided free meals and medical inspections (later treatment) for needy schoolchildren. The 1908 Old Age Pension Act granted weekly pensions to persons aged 70 and over whose annual income was below a certain level, and the National Insurance Act of 1911 established compulsory systems of health insurance (covering all manual workers) and unemployment insurance (covering workers in a limited number of industries). The Liberal welfare reforms provided assistance to the working class that was outside the Poor Law and therefore did not involve 'the stigma of pauperism', and they paved the way for the eventual abolition of the Poor Law.

During the inter-war period the Poor Law served as a residual safety net, assisting those who fell through the cracks of existing social insurance policies. A large share of those on relief, especially in the mid-1920s, were unemployed workers who either did not qualify for unemployment benefits or had exhausted their benefits. The Local Government Act of 1929 abolished the Poor Law unions, and transferred the administration of poor relief to the counties and county boroughs. Finally, from 1945 to 1948, Parliament adopted a series of laws that together formed the basis for the welfare state, and made the Poor Law redundant. The National Assistance Act of 1948 officially repealed all existing Poor Law legislation.

See Also

- ▶ Poor Law, Old
- ▶ Poverty
- ▶ Poverty Alleviation Programmes
- ▶ Social Insurance
- ▶ Social Insurance and Public Policy
- ▶ Welfare State

Bibliography

Boot, H.M. 1990. Unemployment and poor law relief in Manchester, 1845–50. *Social History* 15: 217–228.

Booth, C. 1894. *The aged poor in England and Wales*. London: Macmillan.

Boyer, G.R. 1990. *An economic history of the english poor law, 1750–1850*. Cambridge: Cambridge University Press.

Boyer, G.R. 2004. The evolution of unemployment relief in Great Britain. *Journal of Interdisciplinary History* 34: 393–433.

Brundage, A. 2002. *The english poor laws, 1700–1930*. New York: Palgrave.

Crowther, M.A. 1981. *The workhouse system, 1834–1929: The history of an english social institution*. London: Batsford.

Digby, A. 1978. *Pauper palaces*. London: Routledge & Kegan Paul.

Englander, D. 1998. *Poverty and poor law reform in Britain: From Chadwick to Booth, 1834–1914*. London: Addison Wesley Longman.

Fraser, D. 1976. *The new poor law in the nineteenth century*. London: Macmillan.

Humphreys, R. 1995. *Sin, organized charity, and the poor law in Victorian England*. New York: St Martin's.

Humphreys, R. 2001. *Poor relief and charity, 1869–1945: The London charity organization society*. New York: Palgrave.

Johnson, P. 1985. *Saving and spending: The working-class economy in Britain, 1870–1939*. Oxford: Clarendon Press.

Kidd, A. 1999. *State, society and the poor in nineteenth-century England*. London: Macmillan.

Lees, L.H. 1998. *The solidarities of strangers: The english poor laws and the people, 1770–1948*. Cambridge: Cambridge University Press.

MacKinnon, M. 1986. Poor law policy, unemployment, and pauperism. *Explorations in Economic History* 23: 299–336.

MacKinnon, M. 1987. English poor law policy and the crusade against outrelief. *Journal of Economic History* 47: 603–625.

Rose, M.E. 1970. The new poor law in an industrial area. In *The industrial revolution*, ed. R.M. Hartwell. Oxford: Oxford University Press.

Rose, M.E. 1981. The crisis of poor relief in England, 1860–1890. In *The emergence of the welfare state in Britain and Germany, 1850–1950*, ed. W.J. Mommsen. London: Croom Helm.

Rose, M.E. 1985. *The poor and the city: The english poor law in its urban context*. Leicester: Leicester University Press.

Thomson, D. 1984. The decline of social security: Falling state support for the elderly since early Victorian times. *Ageing and Society* 4: 451–482.
Webb, S., and B. Webb. 1929. *English poor law history. Part II: the last hundred years*, vol. 2. London: Longmans.
Williams, K. 1981. *From pauperism to poverty.* London: Routledge.

Poor Law, old

George R. Boyer

Abstract

The Old Poor Law was the system of local public assistance that existed in England and Wales from 1597 until 1834. It provided an important safety net for labouring households that were unable to protect themselves against income loss, assisting the elderly, widows, children, the sick, and the unemployed. Relief expenditures increased sharply from 1750 to 1820, as did the share of paupers who were adult able-bodied males. Parliament responded to the increase in spending with the Poor Law Amendment Act (1834), which recommended that poor relief be granted to able-bodied males and their families only in workhouses.

Keywords

Chadwick, E.; Enclosures; Labour mobility; Malthus, T.; Old Poor Law; Poor Law; Amendment Act 1834; Poor rate; Poverty alleviation programmes; Senior, N; Settlement Law; Smith, A.; Speenhamland system

JEL Classifications

N4

The Old Poor Law was the system of public assistance in England and Wales from the Tudor era through the passage of the Poor Law Amendment Act in 1834. Parliamentary acts of 1597–98 and 1601 (43 Eliz. I c. 2) established a compulsory system of poor relief administered and financed at the parish (local) level. Overseers of the poor assessed a compulsory property tax, known as the poor rate, to assist those within the parish ‘having no means to maintain’ themselves. The overseers were to put the able-bodied poor to work, give apprenticeships to poor children, and provide ‘competent sums of money’ to relieve the aged or non-able-bodied.

The Elizabethan Poor Law was an attempt by Parliament both to prevent starvation and to ensure public order. It was adopted in response to a sharp deterioration in workers’ living standards in the 16th century, combined with a decline in traditional forms of charitable assistance. The dissolution of the monasteries, religious guilds, almshouses, and hospitals under Henry VIII had eliminated many of the traditional sources of charity for the poor.

The Settlement Act of 1662 stated that individuals were guaranteed relief only in their parish of settlement (typically their parish of birth). The act gave parishes the right to remove within 40 days of arrival any newcomer deemed ‘likely to be chargeable’ as well as any non-settled applicant for relief. Adam Smith believed that the Settlement Law put a serious brake on labour mobility, but available evidence suggests that parishes used it selectively, to keep out economically undesirable migrants such as single women, older workers and men with large families. The Removal Act of 1795 amended the Settlement Law so that no non-settled person could be removed from a parish unless he or she applied for relief.

The Old Poor Law constituted ‘a welfare state in miniature’, relieving the elderly, widows, children, the sick, the disabled, and the unemployed and underemployed (Blaug 1964). It provided an important safety net for labouring households who were unable to accumulate enough savings to protect themselves against income loss. While only a small share of the labouring population received relief at any point in time, the life-cycle nature of poverty meant that a much larger share required Poor Law assistance during their lifetimes. Slack (1990) estimates that in the late 18th century one-fifth or more of the inhabitants of a typical parish received poor relief over a five-year period.

In years of exceptionally high food prices, the share on relief could exceed 25 per cent.

During the 17th century the bulk of relief recipients were elderly, orphans, or widows with young children. In many parishes a majority of those collecting regular weekly pensions were aged 60 or older. Female pensioners far outnumbered males. On average, the payment of weekly pensions made up about two-thirds of relief spending; the remainder went to casual benefits, often to able-bodied males in need of short-term relief because of sickness or unemployment.

Growth in Relief Expenditures, 1750–1820

The 18th century witnessed an explosion in relief expenditures, as can be seen from Table 1. Real per capita expenditures increased by 80 per cent from 1696 to 1748–50, more than doubled from 1750 to 1803, and then remained at a high level until the Poor Law was amended in 1834. Relief expenditures increased from 0.8 percent of GDP in 1696 to a peak of 2.7 per cent of GDP in 1818–20. The demographic characteristics of the 'pauper host' changed considerably in the late 18th and early 19th centuries, especially in the rural south and east of England. There was a sharp increase in numbers receiving casual benefits, as opposed to regular weekly pensions. The share of paupers aged 20–59 increased significantly, and the share aged 60 and over declined. Finally, the share of relief recipients in the south and east who were male increased from about a third in 1760 to nearly two-thirds in 1820. In the north and west there also were shifts toward prime-age males and casual relief, but the magnitude of these changes was far smaller than elsewhere.

Poor Law, old, Table 1 Poor relief expenditures 1696–1841

Year	Expenditures on relief (£1,000s)	Real expenditure per capita, 1803 = 100	Expenditures as % of GDP
1696	400	24.9	
1748–50	690	45.8	1.0
1776	1,530	64.0	1.6
1783–85	2,004	75.6	1.8
1803	4,268	100.0	2.2
1813	6,656	91.8	2.6
1818	7,871	116.8	
1821	6,959	113.6	2.7
1826	5,929	91.8	
1831	6,799	107.9	2.0
1836	4,718	81.1	
1841	4,761	61.8	1.1

Note: Relief expenditure data are for the year ended on 25 March. In calculations of real per capita expenditures, cost of living and population data for the previous year were used

Sources: Data in column 1: Slack (1990: 30) and Mitchell (1988: 605). Data in column 2: author's calculations. Data in column 3: Lindert (1998: 114)

What caused the sharp increase in the number of able-bodied males on relief? In the second half of the 18th century, a large share of rural households in southern England suffered significant declines in real income, resulting from the combination of a decline in agricultural labourers' real wage rates, an increase in seasonal unemployment, a decline in employment opportunities for women and children in cottage industry and, in some villages, the loss of access to land for growing food, grazing animals, and gathering fuel (common rights) as a result of enclosures. The situation was different in the north and midlands, where real wages of day labourers in agriculture increased from 1770 to 1820. Moreover, while some areas experienced a decline in cottage industry, in Lancashire and the West Riding of Yorkshire the concentration of textile production led to increased employment opportunities for women and children.

Forms of Relief and Regional Differences in Relief Spending

Relief for able-bodied males and their families took various forms, the most important of which were: allowances-in-aid-of-wages (the so-called Speenhamland system), child allowances for labourers with large families, and payments to seasonally unemployed agricultural labourers. Under the allowance system, a household head (whether employed or unemployed) was guaranteed a minimum weekly income, the level of which was determined by the price of bread and by the size of his or her family. The most famous allowance scale was that adopted by Berkshire magistrates at Speenhamland in May 1795. Such scales typically were instituted during years of high food prices, such as 1795–6 and 1800–1, and removed when prices declined. Child allowance payments were widespread in the rural south, which suggests that labourers' wages were too low to support large families. The typical parish paid a small weekly sum to labourers with four or more children under age 10 or 12. Seasonal unemployment had been a problem for agricultural labourers long before 1750, but the extent of seasonality increased in the second half of the 18th century as farmers in southern and eastern England responded to the sharp increase in grain prices by increasing their specialization in grain production. The increase in seasonal unemployment, combined with the decline in other sources of income, forced many agricultural labourers to apply for poor relief during the winter.

Table 2 reports data for 15 counties located throughout England on per capita relief expenditures for the years ending in March 1803, 1812 and 1831, and on relief recipients in 1802–3. Per capita expenditures were higher on average in agricultural counties than in more industrial counties, and were especially high in the grain-producing south-eastern counties. The share of the population receiving poor relief in 1802–3 varied significantly across counties, being 15–23 per cent in the grain-producing south and less than 10 per cent in the north. The demographic characteristics of those relieved also differed across regions. The share of relief recipients who were elderly or disabled was higher in the north and west than in the south, while the share who were able-bodied was higher in the south-east than

Poor Law, old, Table 2 County-level poor relief data, 1802–1831

Country	Per capita relief spending (shillings per year) 1802–3	Per capita relief spending (shillings per year) 1812	Per capita relief spending (shillings per year) 1831	% of population relieved 1802–3	Share of recipients over 60, or disabled 1802–3
North					
Durham	6.5	9.9	6.8	9.3	22.8
Northumberland	6.7	7.9	6.3	8.8	32.2
Lancashire	4.4	7.4	4.4	6.7	15.0
West Riding	6.5	9.9	5.6	9.3	18.1
Midlands					
Stafford	6.9	8.5	6.5	9.1	17.2
Nottingham	6.3	10.8	6.5	6.8	17.3
Warwick	11.3	13.3	9.6	13.3	13.7
South-east					
Oxford	16.2	24.8	16.9	19.4	13.2
Berkshire	15.1	27.1	15.8	20.0	12.7
Essex	12.1	24.6	17.2	16.4	12.7
Suffolk	11.4	19.3	18.3	16.6	11.4
Sussex	22.6	33.1	19.3	22.6	8.7
South-west					
Devon	7.3	11.4	9.0	12.3	23.1
Somerset	8.9	12.3	8.8	12.0	20.8
Cornwall	5.8	9.4	6.7	6.6	31.0
England and Wales	8.9	12.8	10.1	11.4	16.0

Sources: Data for columns 1–3: Blaug (1963: 178–9). Data for columns 4–5: *Abstract of Returns relative to the Expense and Maintenance of the Poor*, H.C. 175 (1803–4), xiii elsewhere. These regional differences in relief expenditures and numbers on relief largely were caused by differences in economic circumstances; poverty was more of a problem in the agricultural south and east than it was in the pastoral southwest or in the more industrial north (Blaug 1963; Boyer 1990). Recently, King (2000, pp. 267–8) has argued that the regional differences in poor relief were determined by 'very different welfare cultures on the part of both the poor and the poor law administrators'.

Political Economy of Poor Relief

From 1795 to 1834 relief expenditures as a share of national product were significantly higher in England than on the European continent. However, differences in spending between England and the continent were relatively small before 1795 and after 1834 (Lindert 1998). The increase in relief spending in late 18th and early 19th century England overstates the increase in poverty, because it was partly a result of politically dominant farmers taking advantage of the poor relief system to shift some of their labour costs onto other taxpayers (Boyer 1990). Most rural parish vestries were dominated by labour-hiring farmers as a result of the system of plural voting introduced by Gilbert's Act in 1782 and extended in 1818 by the Parish Vestry Act, which gave large property holders (typically labour-hiring farmers) up to six votes in local elections. Relief expenditures were financed by a tax levied on all parishioners whose property value exceeded some minimum level. A typical rural parish's taxpayers can be divided into two groups: labour-hiring farmers and non-labour- hiring taxpayers (tithe recipients, family farmers, shopkeepers, and artisans).

In grain-producing areas, where there were large seasonal variations in the demand for labour, labour-hiring farmers anxious to secure an adequate peak season labour force were able to reduce costs by laying off unneeded workers during slack seasons and having them collect poor relief. Tithe recipients and other non-labour-hiring taxpayers paid part of the relief benefits that went to seasonally unemployed labourers. Thus, some share of the increase in relief spending in the early 19th century represented a subsidy to labour-hiring farmers rather than a transfer from farmers and other taxpayers to agricultural labourers and their families. In pasture farming areas, where the demand for labour was fairly constant over the year, it was not in farmers' interests to shed labour during the winter, and the number of able-bodied labourers receiving casual relief was smaller.

Reform of the Poor Law

The sharp increase in relief spending after 1780 sparked a major debate on the Poor Laws. Most participants in the debate were critical of the granting of outdoor relief to able-bodied males, on the grounds that such aid created serious work disincentives. Among the sharpest critics was Thomas Malthus, who argued in *An Essay on the Principle of Population* (1798, pp. 40–1) that the Poor Laws, by guaranteeing parish assistance to able-bodied labourers, 'diminish both the power and the will to save among the common people, and thus . . . weaken one of the strongest incentives to sobriety and industry, and consequently to happiness'.

In 1832 the government appointed the Royal Commission to Investigate the Poor Laws to examine the operation of the Poor Law and suggest methods for improving the administration of relief. The commission's report (1834, pp. 261–3), written by economists Nassau Senior and Edwin Chadwick, called for sweeping reforms, including the abolition of outdoor relief for the able-bodied and their families. The report urged the adoption of a policy of 'less eligibility' whereby the condition of paupers would be worse than that of the lowest-paid independent labourers. To achieve this, Senior and Chadwick recommended that relief should be granted to able-bodied labourers and their families only in well-regulated workhouses; they predicted that the use of workhouses would restore the industry and 'frugal habits' of the poor, and improve their 'moral and social condition'.

The era of the Old Poor Law ended with the adoption of the Poor Law Amendment Act of 1834, which implemented many of the report's recommendations.

See Also

- ▶ Eden, Frederick Morton (1766–1809)
- ▶ Malthus, Thomas Robert (1766–1834)
- ▶ Poor Law, New
- ▶ Poverty Alleviation Programmes
- ▶ Senior, Nassau William (1790–1864)
- ▶ Welfare State

Bibliography

Blaug, M. 1963. The myth of the Old Poor Law and the making of the New. *Journal of Economic History* 23: 151–184.

Blaug, M. 1964. The Poor Law report re-examined. *Journal of Economic History* 24: 229–245.

Boyer, G. 1990. *An economic history of the English Poor Law, 1750–1850*. Cambridge: Cambridge University Press.

King, S. 2000. *Poverty and welfare in England, 1700–1850: A regional perspective*. Manchester: Manchester University Press.

Lees, L. 1998. *The solidarities of strangers: The English Poor Laws and the people, 1770–1948*. Cambridge, MA: Cambridge University Press.

Lindert, P. 1998. Poor relief before the welfare state: Britain versus the Continent, 1780–1880. *European Review of Economic History* 2: 101–140.

Malthus, T. 1798. *An essay on the principle of population*. Oxford: Oxford University Press. 1999.

Mitchell, B. 1988. *British historical statistics*. Cambridge: Cambridge University Press.

Royal Commission to Investigate the Poor Laws. 1834. *Report on the administration and practical operation of the Poor Laws*. London: B. Fellowes.

Slack, P. 1990. *The English Poor Law, 1531–1782*. London: Macmillan.

Webb, S., and B. Webb. 1927. *English Poor Law history. Part I: The Old Poor Law*. London: Longmans.

Williams, K. 1981. *From pauperism to poverty*. London: Routledge.

Population Ageing

David N. Weil

Abstract

Population ageing is primarily the result of past declines in fertility, which produced a decades-long period in which the ratio of dependents to working-age adults was reduced. Rising old-age dependency in many countries represents the inevitable passing of this 'demographic dividend'. Societies use three methods to transfer resources to people in dependent age groups: government, family, and personal saving. In developed countries, families are predominant in supporting children, while government is the main source of support for the elderly. The most important means by which ageing will affect aggregate output is the distortion from taxes to fund public pensions.

Keywords

Bequests; Capital accumulation; Capital–labour ratio; Demographic dividend; Demography; Dependency; Family transfers; Fertility in developed countries; Fertility in developing countries; Generational crowding; India; International capital flows; Japan; Labour supply; Mortality; Old-age pensions; Optimal saving; Payroll taxes; Population ageing; Retirement; Saving; Social Security in the United States

JEL Classifications

J11; J10; H55

Population ageing is the shift in the distribution of a country's population towards older ages. An increase in the population's mean or median age, a decline in the fraction of the population composed of children, or a rise in the fraction of the population that is elderly are all aspects of population ageing.

Population ageing is occurring in most parts of the world, but is most advanced in the richest countries. Among the countries currently classified by the United Nations as more developed (which had a population of 1.2 billion in 2005), the median age of the population rose from 29.0 in 1950 to 37.3 in 2000, and is forecast to rise to 45.5 by 2050. The corresponding figures for the world as a whole are 23.9 for 1950, 26.8 for 2000, and 37.8 for 2050. In Japan, one of the fastest-ageing countries in the world, in 1950 there were 9.3 people younger than 20 for every person older than 65. By the year 2025, the ratio is forecast to be 0.59 people younger than 20 for every person older than 65 (United Nations 2004).

The sources of population ageing lie in two demographic phenomena: rising life expectancy and declining fertility. An increase in longevity raises the average age of the population by raising the number of years that each person is old relative to number of years in which he is young. A decline in fertility increases the average age of the population by changing the balance of people born recently (the young) to people born further in the past (the old). Of these two forces, it is declining fertility that is the dominant contributor to population ageing in the world today (Weil 1997). More specifically, it is the large decline in the total fertility rate since the 1950s that is primarily responsible for the population ageing that is taking place in the world's most developed countries. Because many developing countries are going through faster fertility transitions, they will experience even faster population ageing than the currently developed countries in the future.

While the economic underpinnings of the demographic processes that cause population ageing – in particular declining fertility – are interesting topics in and of themselves, this article instead concentrates on how ageing affects the economy.

The Economic Effects of Population Ageing

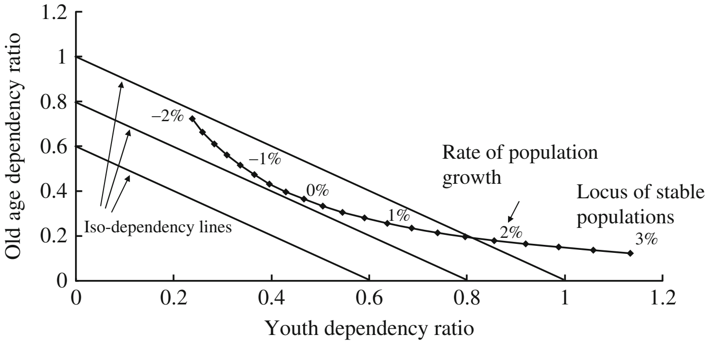
Population ageing has economic effects whenever some economic interaction (the sale of a good or service, the provision of a government benefit, and so on) brings together people whose participation is a function of their age. In such a situation, a change in the relative size of two age groups will require a change in behaviour by members of at least one group. For example, babies demand strollers, which are produced by working-age adults. Thus, a reduction in the ratio of babies to adults will mean more strollers per baby, fewer adults working in stroller production, or both. The changes in behaviour required to restore equilibrium in the face of demographic change are induced through either prices or institutions. If individuals on at least one side of the transaction respond elastically to price changes (as would be the case in getting working-age adults to move from stroller manufacture into the wheelchair business), then the effects of population ageing will be little worth commenting on. But when individuals on both sides of the interaction are not easily induced to change their behaviour, the economic effects of population ageing will be dramatic. Old-age pensions, child rearing, and the combining of old people's capital with young people's labour are all cases where a change in the relative numbers on either side of the equation will have important effects.

The simplest analysis of the economic effects of population ageing starts with the notion of age-based dependency: people of some ages produce less than they consume, and are dependent on the rest of society for their support. Consider a division of the population into three age groups: working age adults, dependent youths, and dependent elderly. We temporarily ignore the question of how resources are transferred from working-age adult to dependent children and elderly. For simplicity, we assume that people of all ages have the same consumption, although the analysis can easily be extended to allow for age-varying consumption needs (see Weil 1999). Finally, we assume that output is produced solely by the labour of working-age adults, with no additional factors of production such as capital.

P

The consumption possibilities of our idealized society can be analysed in a diagram like Fig. 1. The horizontal axis plots youth dependency ratio (population aged 0–19 divided by population aged 20–64); the vertical axis plots the old-age dependency ratio (population aged 65+ divided by population aged 20–64). A society's demographic structure is represented by a point in this space. For example, a newly planted colony might be represented by a point in the lower left-hand corner, with youth and old age dependency ratios of zero. For a normal society, however, the demographic processes of ageing, mortality and fertility will determine predictable movements of the age structure through the space of Fig. 1. A set of points of particular interest are what demographers call stable populations. These are populations in which age-specific mortality and

Population Ageing, Fig. 1 Stable populations and iso-dependency lines (*Source:* Author's calculations)



fertility rates have been constant for sufficient time that the relative number of people of each age is constant. Fig. 1 shows a typical locus of stable populations, generated using age-specific mortality rates for the United States in 2000 and varying the level of fertility. The labels show the population growth rate consistent with different points on the locus of stable populations.

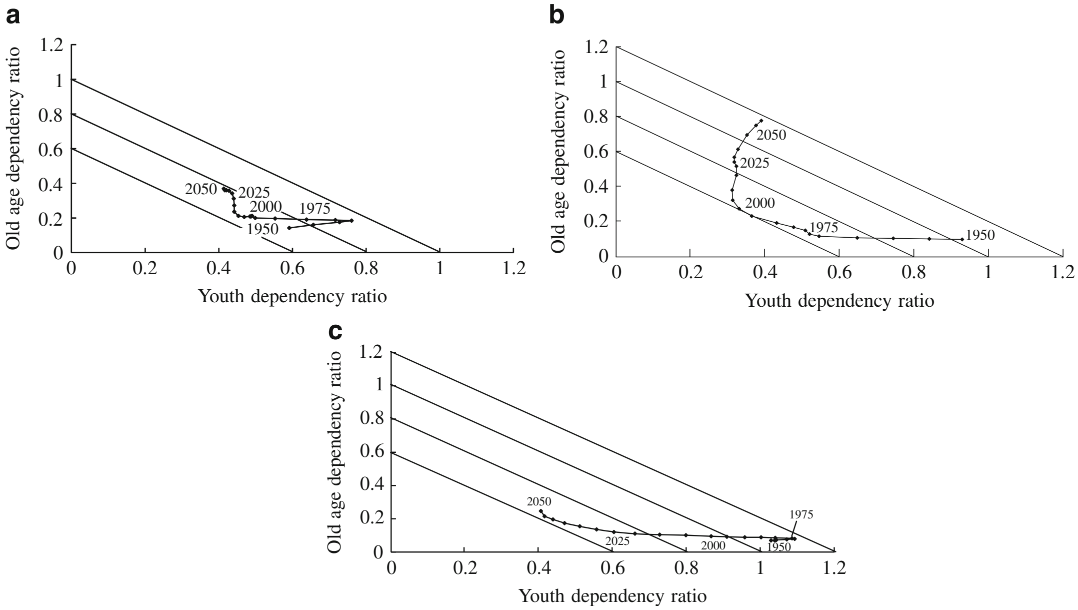
We can also represent in this space the effect of demographic structure on the consumption possibilities of the society through a series of iso-dependency lines. These are lines along which the sum of youth and old-age dependency is constant – in other words, combinations of old-age and youth dependency that yield constant levels of consumption per capita. Iso-dependency lines closer to the origin represent age structures which allow for higher consumption per capita. The tangency between the locus of stable populations and an iso-dependency line shows the stable population with the lowest dependency ratio.

Reductions in fertility will lead to clockwise movements of the point representing a country's demographic structure through the space of Fig. 1. Falling fertility reduces the youth dependency ratio immediately, and only raises the old age dependency ratio with a lag of several decades. For this reason, a country experiencing fertility transition will be able to move temporarily below the locus of stable populations.

FIgure 2a–c show data on population age structure for the United States, Japan, and India over the period 1950–2050. In all cases, the clockwise motion and period of temporarily low dependency due to fertility transition are visible, although the countries differ in how far along they are and how severe the process of ageing is forecast to be. In Japan, the total dependency ratio (youth plus old age) will rise from 0.64 to 1.17 over the period 2005–50, implying, *ceteris paribus,* that GDP per capita will grow 0.6 per cent per year more slowly than GDP per worker (see Weil 2005, ch. 5, for details of this calculation). By contrast, India, like many developing countries, is in the process of receiving a large 'demographic dividend' from reduced fertility (Bloom and Williamson 1998).

The lesson from this analysis of dependency is that, from the point of view of society as a whole, the period of rapid increase in old-age dependency that is in store for the world's richest countries is to a large extent simply the passing of the transitory benefit derived from a decrease in fertility. A second lesson is that any change in fertility that will in the long run undo the effects of population ageing will, in the short run, lead to an *increase* in total dependency by moving the point representing age structure above and to the right of the locus of stable populations.

The model discussed above ignores the means by which dependent members of society are supported. In practice, there are three mechanisms by which this takes place: through their own past savings; through institutions (primarily the government) that transfer resources between unrelated people of different ages; and through their own families. Lee (2002) refers to these various means by which resources are transferred among age groups as a 'reallocation system'. We shall see that the nature of the reallocation system affects the overall burden of ageing as well as the distribution of that burden.



Population Ageing, Fig. 2 (**a**) Demographic dynamics in the United States (Source: United Nations) (**b**) Demographic dynamics in Japan (Source: United Nations) (**c**) Demographic dynamics in India (Source: United Nations)

Ageing, Savings and Capital

Capital is important in analyses of population ageing for two reasons. First, accumulation of capital allows either individuals or society as a whole to break the temporal link between production and consumption: an individual, for example, can save some of her wages when she is working, and then use the accumulated capital to fund consumption during retirement. Second, as a factor of production complementary to labour, capital helps determine the quantity of output to be divided among workers and dependents. Analyses of ageing and capital accumulation proceed down both normative and positive channels.

The normative approach asks how society should respond to a looming change in demographics. Although there is in practice no social planner who makes saving decisions for society as a whole, the solution to the social planner's problem can inform the response of a government that influences national saving through fiscal policy and tax incentives. Common sense would suggest that a country that is undergoing population ageing should 'save for its old age', that is, accumulate extra capital during the period of low dependency in order to maintain a smooth path of consumption into the period of high dependency. As stressed by Cutler et al. (1990), however, there is a countervailing effect: population ageing due to lower fertility implies that the working-age population will grow more slowly, reducing the amount of investment required to supply new workers with capital. The flip side of this decrease in required investment is that, if a country did attempt to save sufficient capital to smooth consumption in the face of ageing, the result would be a rise in the capital–labour ratio, lowering the return on capital, which would lead households (or a social planner) to want to raise consumption. Elmendorf and Sheiner (2000) calculate that an optimizing social planner would want to make relatively small changes in saving rates in response to the population ageing currently forecast in the United States.

A positive alternative to the social planner approach is to consider the equilibrium of an economy in which consumers make privately optimal saving decisions taking as given the expected paths of interest rates and wages as well as taxes and government benefits. Forecasting the effects of demographic change on output or capital per worker, interest rates, and so on requires a fully articulated, rational expectations general equilibrium model. Kotlikoff et al. (2007) use such a model to analyse demographic change in the United States, under the assumption that the Social Security benefit regime does not change, and that payroll taxes adjust accordingly. They find that the capital deepening that would normally accompany a shift of the population into its peak asset-holding years is undone by rising payroll taxes. They forecast 'capital shallowing' that will raise the real return on capital by one percentage point by 2030 and a further two percentage points over the rest of the 21st century, as well as a dramatic slowing of real wage growth.

Rather than fitting an optimizing model of saving, another approach is to look empirically at the age pattern of actual behaviour. Poterba (2005) shows that individual net worth follows a hump-shaped path over the course of the lifetime, peaking between ages 65 and 69. Unlike the classical life-cycle model, however, the decline in average net worth is relatively slow, so that average net worth at death is significant. This life-cycle pattern of asset accumulation in turn implies that shifts in demography will shift asset demands and potentially asset prices. In particular, the movement of the baby-boom generation into its high accumulation years was widely cited as a potential explanation for the run-up in stock prices during the last decades of the 20th century (Abel 2003). Similarly, some analysts have suggested that, as the balance between age groups actively accumulating and running down wealth shifts in the period after 2010, there will be a corresponding meltdown of asset prices. However, Poterba (2005) finds little evidence of demographic effects on asset returns in time series data from the United States, Canada and the United Kingdom. Lim and Weil (2003) show that in a forward-looking asset pricing model it would require an unreasonably large adjustment cost for capital to produce a large asset price meltdown in response to projected population ageing. The shift in population towards the elderly will also lead to a significant increase in the flow of bequests relative to either income or wealth of the younger generation; Weil (1994) argues that this increased flow of bequests will reduce the saving of the receiving generation.

The above discussion implicitly considered the case of an economy closed to international capital flows. In an open economy, the mismatch between the demographically induced demand for asset holding and the capital requirements of the labour force can be channelled into capital flows abroad. For example, a country like India, where the working-age population is forecast to grow an annual rate of 1.8 per cent per year between 2000 and 2025, would be a natural recipient of investment from Japan, where the working-age population will shrink at an annual rate of 0.6 per cent per year over this period. In practice, however, net financial flows among countries tend to be far smaller than a model of perfectly open capital markets would imply, and movements in current accounts seem to bear little resemblance to those predicted by demographic change (see Brooks 2003).

Ageing and Government

In the developed countries that are ageing most rapidly, government transfer programmes are a major source of support for dependent elderly. In Germany, for example, transfers net of taxes and inclusive of public health benefits make up 65 per cent of the income of people aged 65 and older (Burtless 2006).

Correspondingly, one of the most important functions of government is transferring resources to elderly people. In 2005, US federal outlays were 18.9 per cent of GDP. Almost 60 per cent of that amount was spent on direct transfers attributable to specific age groups (Medicare for the elderly, unemployment insurance for working age, and so on). Of such transfers, 58 per cent (6.5 per cent of GDP) was directed toward those aged 65 and older. On a per person basis, the elderly received close to eight dollars in direct transfers for every dollar of transfers received by working-aged persons. In sharp contrast, children received just 35 cents per dollar of transfers awarded to workers. Assuming constant transfers per person by age group, a shift of ten per cent of the population out of the workforce and into retirement would increase federal transfer outlays by 4.7 per cent of GDP (calculations based on data underlying Gokhale and Smetters 2006). In addition to raising spending, population ageing also reduces government revenue. Putting these tax and spending effects together, Burtless (2006) calculates that the effect of population ageing would raise the tax rate required to pay for government transfers on a PAYGO basis in from 16 per cent in 2000 to 21 per cent in 2030 in the United States. In Germany, where transfers are larger and ageing more extreme, the increase in the tax rate would be from 28 per cent to 40 per cent. In the United States, the effect of ageing on the government budget is greatly exacerbated by the fact that the *price* of health care for the elderly is rising at the same time as the fraction of the population that is elderly (Elmendorf and Sheiner 2000).

One important way in which transfers to dependents (either children or elderly) that are channelled through the government differ from those mediated by either the family or through one's own saving is in how workers perceive the benefits resulting from their forgone consumption. People give money and other resources to their children or aged parents because they care about them. And when people save for their own old age, it is because they care about their future selves. But few people are so altruistic that they value the taxes that are taken from their pay in order to fund transfers to the elderly. For this reason, there is an efficiency loss associated with government support of the elderly that is not present in other forms of transfers to dependent age groups. Prescott (2004) argues that differences in marginal labour tax rates explain large cross-sectional differences and changes over time in labour supply among the G-7 countries. For example, in the early 1990s his calculations show the French average marginal tax rate (inclusive of consumption taxes) being 48 per cent larger than that in the United States; correspondingly, French adults aged 15–64 worked only 68 per cent as many hours as their US counterparts. The large elasticity of labour supply that Prescott estimates implies that deadweight losses will increase dramatically as populations age, as long as government old-age pensions continue to be funded through taxes that are largely divorced from the benefits that the individuals paying them will receive. Thus an economy that could function smoothly with a high level of youth dependency funded through family transfers, or a high level of old-age dependency funded through savings, might collapse if a similar level of old-age dependency were funded through taxes.

Because government transfers are so heavily weighted toward the elderly, the adjustment in government finances required to deal with population ageing will be proportionally much larger than the overall change in consumption in the economy as a whole. Roughly put, ageing is a much bigger problem for the government than for the economy as a whole. Most conceivable reforms in government old-age pensions will represent net losses to cohorts who are near or beyond retirement at the time of reform. Bohn (2005) calculates that, based on current participation rates, the fraction of voters aged 65 and over in the United States will rise from 19.8 per cent to 30.5 per cent between 2003 and 2030; over the same period the age of the median voter will rise from 47 to 52. Thus, as the fiscal strain from population ageing becomes acute it will be increasingly difficult for policymakers to solve their problems by reducing transfers to the elderly.

Ageing and Families

Transfers within families represent the final channel whereby dependents are supported. For the large majority of old people in developed countries, family transfers are the second or third most important source of support, behind their own past savings and/or transfers from the government. This is a relatively new pattern. Prior to the 20th century, the period of old-age dependency was much shorter, government transfers to the elderly were minimal, and cohabitation of elderly with their children was the norm. In the United States, for example, the fraction of elderly widows who lived with their adult children fell from 67 per cent in 1920 to 20 per cent in 1990 (McGarry and

Schoeni, 2002). Only 2.7 per cent of people aged 60 and over in the United States reported support from children as their main source of income in 2001. Even in Japan, where such transfers have traditionally played a much larger role, the fraction of people 60 and over reporting their children as their main source of support fell from 29.8 per cent to 12.0 per cent between 1981 and 2001 (United Nations 2005, Table I.2). In contrast to the elderly, the burden of supporting young dependents lies foremost on their own families. Mason et al. (2005) calculate that 57 per cent of consumption of people under 20 in the United States in 2000 was financed by transfers from family members. Thus, unlike governments, families headed by working-age adults find their budget constraints relaxed by the low fertility that causes population ageing.

An important distinction between support for elderly dependents and support for child dependents concerns the degree of choice that those doing the support enjoy. Working-age adults cannot choose how many siblings they share the burden of supporting their parents with, much less the size of the working-age cohort relative to the elderly population, which determines the level of taxation required to fund public pensions. But working-age adults *can* choose the number of children they produce and support, and their choices about fertility may respond to economic conditions. Of particular interest in the present context, population ageing itself may feed back to affect fertility. The best-known mechanism whereby population age structure affects fertility was identified by Easterlin (1987), who hypothesized that members of large birth cohorts would suffer from labour market crowding, earn wages that are low relative to the standard of living that they had grown up with, and would adjust fertility downward to partially restore their standard of living. The rise in taxes required to fund transfers to the elderly that will result from population ageing could have effects on after-tax income that are as large or larger than those from Easterlin-style generational crowding; thus, ageing could lead to lower fertility and, down the road, even more ageing (Hock and Weil 2006).

See Also

- ▶ Economic Demography
- ▶ Fertility in Developed Countries
- ▶ Retirement
- ▶ Social Security in the United States

Acknowledgment *I am grateful to Jagadeesh Gokhale for helpful comments.*

Bibliography

Abel, A. 2003. The effects of a baby boom on stock prices and capital accumulation in the presence of Social Security. *Econometrica* 71: 551–578.

Bloom, D., and J. Williamson. 1998. Demographic transitions and economic miracles in emerging Asia. *World Bank Economic Review* 12: 419–456.

Bohn, H. 2005. Will social security and medicare remain viable? In *Social security reform: Financial and political issues in international perspective*, ed. R. Brooks and A. Razin. Cambridge, UK: Cambridge University Press.

Brooks, R. 2003. Population ageing and global capital flows in a parallel universe. *IMF Staff Papers* 50(2): 200–221.

Burtless, G. 2006. Cross-national evidence on the burden of age-related public transfers and health benefits. Working Paper No. 2006-6. Center for Retirement Research, Boston College.

Cutler, D., J. Poterba, L. Sheiner, and L. Summers. 1990. An ageing society: Opportunity or challenge? *Brookings Papers on Economic Activity* 1990(1): 1–73.

Easterlin, R. 1987. *Birth and fortune: The impact of numbers on personal welfare*. 2nd ed. Chicago: University of Chicago Press.

Elmendorf, D., and L. Sheiner. 2000. Should America save for its old age? Fiscal policy, population ageing, and national saving. *Journal of Economic Perspectives* 14(3): 57–74.

Gokhale, J., and K. Smetters. 2006. Fiscal and generational imbalances: An update. In *Tax policy and the economy, volume 20*, ed. J. Poterba. Cambridge, MA: MIT Press.

Hock, H., and D. Weil. 2006. *The dynamics of age structure, dependency, and consumption*. Mimeo/Brown University.

Kotlikoff, L., K. Smetters, and J. Walliser. 2007. Mitigating America's demographic dilemma by pre-funding social security. *Journal of Monetary Economics* 54: 247–266.

Lee, R. 2002. A cross-cultural perspective on intergenerational transfers and the economic life cycle. In *Sharing the wealth: Demographic change and economic transfers between generations*, ed. A. Mason and G. Tapinos. Oxford: Oxford University Press.

Lim, K.M., and D. Weil. 2003. The baby boom and the stock market boom. *Scandinavian Journal of Economics* 105(3): 359–378.

Mason, A., R. Lee, A.C. Tung, M.S. Lai, and T. Miller. 2005. *Population ageing and intergenerational transfers: Introducing age into national accounts*. Mimeo/ University of Hawaii.

McGarry, K., and R. Schoeni. 2002. Social security, economic growth, and the rise of independence of elderly widows in the 20th century. *Demography* 37: 221–236.

Poterba, J. 2005. Demographic structure and asset returns. In *Social security reform: Financial and political issues in international perspective*, ed. R. Brooks and A. Razin. Cambridge, UK: Cambridge University Press.

Prescott, E. 2004. Why do Americans work so much more than Europeans? *Federal Reserve Bank of Minneapolis Quarterly Review* 28(1): 2–14.

United Nations. 2004. *World population prospects: The 2004 revision population database*. New York: Population Division, United Nations.

United Nations. 2005. *Living arrangements of older persons around the world*. New York: Department of Economic and Social Affairs, United Nations.

Weil, D. 1994. The saving of the elderly in micro and macro data. *Quarterly Journal of Economics* 109: 55–81.

Weil, D. 1997. The economics of population ageing. In *Handbook of population and family economics*, ed. M. Rosenzweig and O. Stark. Amsterdam: North-Holland.

Weil, D. 1999. Population growth, dependency, and consumption. *American Economic Review* 89: 251–255.

Weil, D. 2005. *Economic growth*. Boston: Addison-Wesley.

Population and Agricultural Growth

James Roumasset

Abstract

Thinking about population as a driver of agricultural development provides insights into induced technical and institutional change, whether it be Ester Boserup's declining fallow period, modern crop varieties, or the horizontal and vertical specialization that arise in labour-intensive agriculture. The non-convexities of research and development, infrastructure investments, and specialization imply that modest population pressure does not necessarily exert downward pressure on wages. As agricultural growth stimulates industrialization, the non-convexities of specialization become ever more compact. The combination of these and the increased demand for human capital, if not inhibited by policy failures, tends to promote a virtuous circle of human progress.

Keywords

Black Death; Capital intensity; Child care; Development economics; Dismal science; Division of labour; Elasticity of substitution; Endogenous growth theory; Endogenous population growth; Exchange labour; Fertility; Governance; Human capital; Immiserizing growth; Indivisibilities; Induced institutional change; Induced technical change; Industrialization; Innovation possibility frontier; Intergenerational neutrality; Kuznets curve; Labour intensification; Labour market contracts; Labour productivity; Learning-by-doing; Leisure; Malthus's theory of population; Mortality; Natural capital; New classical economics; New household economics; Non-convexity; Population and agricultural growth; Population growth; Poverty alleviation; Property rights; Rent seeking; Research and development; Resource depletion; Shadow pricing; Sharecropping; Specialization; Spot markets; Steady state; Subsistence; Sustainability criteria; Sustainable development; Transaction costs

JEL Classifications

Q1

That economics became known as the ‘dismal science’ can largely be attributed to the theory of population and agricultural growth as developed by Malthus and Ricardo, notwithstanding the term's origin in another context. Starting from a point of relatively high wages, for example at the end of the Black Death in Europe, or after some exogenous technological improvement, population increases geometrically. The additional population is assimilated by agricultural growth at the extensive and intensive margins, both of which result in diminishing returns to labour. Extensive growth occurs through the expansion of cultivated land, which Ricardo (1817) presumed to be more distant from or of poorer quality than land already in use. Growth at the intensive margin likewise results in diminishing returns, due to the greater amount of labour and other inputs employed on the fixed quantity of previously cultivated land. As a consequence, Ricardo (1817) and Malthus (1798) theorized that wages would eventually decline towards a subsistence level, where population growth would cease due to ‘positive checks’ such as starvation and disease.

Modern economists still use this dismal theory to explain why growth in levels of living among the working classes was never sustained for long periods until the advent of the Industrial Revolution. Each technological improvement was subsequently ‘eaten up’ by population growth and the subsequent diminishing returns. The belief in this theory is so strong that Lucas (2002, ch. 3) wrote that he could look at a picture of a Korean peasant farm in an unknown century and confidently guess household income. Recent interest in ‘sustainable development’ has augmented resource pessimism. In this view, the conventional Malthusian vicious circle between population growth and poverty is exacerbated by resource depletion and environmental degradation. Expanding numbers of poor people in developing countries put more pressure on limited natural resources and fragile ecosystems, and the falling resource base makes the Malthusian circle even more vicious than with a fixed resource endowment.

Malthus famously argued that unchecked population growth is exponential while food production at best grows linearly, thus implying the inevitability – in the absence of sufficient *preventative* checks – of *positive* checks such as pestilence, plague, famine and war and of subsistence levels of income in the long run. Ironically, food supply has outstripped population growth ever since the publication of Malthus’s *Essay on the Principle of Population.* Technological and institutional change has been more rapid than he envisioned and preventative checks more robust.

Boserup Effects

Boserup (1965, 1981) takes a different tack by taking population growth as the exogenous variable and enquiring into the consequences thereof for agricultural technology and institutional change. I follow Boserup’s lead in most of what follows, eventually returning to a more integrated view. Boserup focused on the effects of physiological population density on an additional intensive margin – the fallow period. As population (and other demand factors) grow, the predominant agricultural system gradually transitions from long to short fallow to annual cropping to multiple cropping. Table 1 describes these systems and

Population and Agricultural Growth, Table 1 Boserup’s frequency of cropping by population density

System	Description of cropping system	Frequency of cropping	Person per km^2	Density
Hunting and gathering	Wild plants, roots, fruits and nuts are gathered	0%	0–2	Very sparse
Forest fallow (w/astoralism)	One or two crops followed by 15–25 years’ fallow	0–10%	1–4	Very sparse
Bush fallow (w/pastoralism)	Two or more crops followed by 8–10 years’ fallow	10–40%	4–64	Sparse to Medium
Short fallow (w/domestic animals)	One or two crops followed by one or	40–80%	16–64	Medium
	two years’ fallow			
Annual cropping (w/intensive animal husbandry)	One crop each year with only a few months’ fallow	80–100%	64–256	Dense
Multi-cropping	Two or more crops in the same fields each year without any fallow	200–300%	$ 256	Very dense

Source: Boserup (1981, pp. 9, 19 and 23) illustrates the rough correspondence between the frequency of cropping and population density in less developed economies. Other authors have extended the correlation between population density and cropping frequency to European countries, both over time and country.

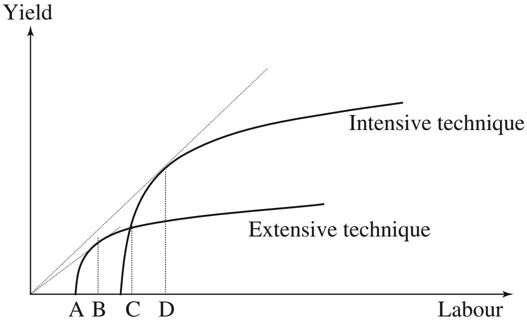
Boserup's insight can be partly understood from the perspective of *induced technical change* (Ahmad 1966). Absent industrial growth, population pressure makes land increasingly scarce relative to labour, thus inducing land-saving technical change. In the era of modern economic growth, the same tendency would influence whether capital was used to save labour or land. This was exemplified by labour- abundant Japan developing land-saving biological innovations and the United States developing labour-saving mechanical innovations (Hayami-Ruttan 1985). As represented with standard neoclassical analysis, however, induced innovation simply increases the elasticity of factor substitution (especially between land and labour). In the *very* long run, that is, allowing for induced technical change, the elasticity of substitution, such as between land and labour, is higher than without technical change.

Similarly, decreasing the fallow period allows the marginal product to decline more slowly than otherwise. For example, suppose that 100 workers cultivate 100 hectares with a 50 per cent cropping frequency (short fallow) and that the population doubles. Even though the additional labour can be productively employed, for instance by better weeding and more thorough land preparation, the marginal product of labour will suffer a large decline if the cropping frequency remains unchanged (perhaps by a half or more). By switching to annual cropping, however, it may be possible to accommodate the additional labour with only a small decline in its marginal product, even in the steady state. The optimal solution involves some conservation of soil fertility over time, for example through the use of animal manure and crop rotation (Barrett 1991).

Boserup contends that it is even possible that population pressure increases the productivity of agricultural labour. More intense farming systems require more fixed costs. For example, forest fallow systems require minimal land preparation. The slash and burn method leaves the land both fertile and weed-free. In the tropical African context that she describes, however, once the land has been burned and cropped, it is taken over by grasses and is no longer suitable for slash and burn agriculture until 20 or more years later, when the forest has returned. Consequently, land preparation requires time-intensive ploughing. Because of these fixed costs, the average product of labour rises over some range.

Other investments associated with intensification, such as irrigation and terracing, similarly increase labour productivity. This is illustrated in Fig. 1. Once population has reached point C, the average product of the extensive and intensive techniques is equalized and it becomes worthwhile to switch to the intensive method. As labour increases beyond C, the average product rises until D, where diminishing returns just offset the gains from spreading the fixed costs, and average product begins to decline. In this sense, population eventually overcomes the transitory gains from switching techniques and causes productivity to fall.

Innovation-through-intensification, as portrayed in Fig. 1, does not require invention. It is as if new techniques are taken 'off the shelf' when they are warranted by increased land scarcity. Genuinely new technology, developed through invention or imported from other areas, may provide additional positive effects. The same population increase that warrants the fixed cost of intensification also



Population and Agricultural Growth, Fig. 1 Average product of labour under different farming techniques (Source: Adapted from Krautkraemer (1994)) warrants increased expenditures on experimentation and research. This research shifts the *innovation possibility frontier* (IPC) between land and labour inwards. In modern settings, R&D becomes an important source of productivity growth.

For example, the high-yielding, or modern, wheat and rice varieties (MVs) developed in the 1960s were in large part induced by population pressure on increasingly scarce land. In the extensive phase of agricultural development, cultivated hectarage is increasing. Eventually, cultivated area reaches a maximum and declines as towns and industrial areas encroach on agricultural land. At this point, land scarcity is exacerbated by both rising food demand and falling land supply, and intensification accelerates.

One of the effects of intensification is to increase the demand for land-saving technology. According to the 'political Boserup effect' (Evenson 2004), increasing population densities induce countries to invest more in the genetic improvement of both crops and animals. By first characterizing existing technology by the unit requirements of land, labour, and capital, optimal investment by a country in new technology can be described by the amount of research and its factor-saving bias. In one version of this theory, a given research expenditure allows a country to pick any point on the IPC, the envelope of all unit isoquants in the land–labour plane, that said research expenditure affords. If it is assumed that the IPC shifts in a neutral fashion towards the ultimate IPC, wherein the marginal benefit of research is zero, then the factor-saving bias is in accordance with changes in relative factor prices. For example, if population growth results in a decrease in the wage rate and an increase in the land rental rate, both relative to the price of capital, then technical change will be land-saving and labour-using relative to capital (Binswanger and Ruttan 1978, chs 2 and 4).

In as much as the IPC shifts in a non-neutral fashion, however, these results will be modified. It is natural to assume, for example, that technical change is inherently capital-using, that the unit isoquant (net of capital costs) can be shifted inward more cheaply by increasing capital per unit of output than by increasing labour or land. Moreover, it may be that inventing technology that uses capital to save labour is cheaper than technology that saves on land. This may explain why the modern rice and wheat varieties have been found to be mildly labour-saving, in addition to being land-saving and capital-using (fertilizer responsive), even though their demand was created by falling wages relative to land rents. But even though labour per unit of output fell, output per hectare increased enough such that MVs had a positive effect on wages (for example, Evenson 1982). Overall, MVs have had a beneficial effect on poverty reduction by decreasing food prices and increasing wages relative to what they would have otherwise been given population growth and labour demand in other sectors.

Boserup's other 'secondary effects' of population growth may also cause productivity to rise, even in the absence of agricultural research. Among these are property rights, work habits, division of labour, education, and the infrastructure for transport and communication. Changing property rights exemplifies how institutions can change in response to population pressure and other changes in factor scarcities. This insight led to the theory of *induced institutional change* as a complement of the theory of induced technical change. For example, as population pressure increased the demand for land-saving investments, private property sometimes emerged as a more efficient substitute for top-down land management by community leaders or feudal lords (see, for example, North and Thomas 1973). Indeed, the first legal enforcement of the early English enclosures was effected by the Statute of Merton (1235), which noted the need to improve the land in order to generate greater rent. The subsequent waves of English enclosures beginning before the 17th and 19th centuries also appear to have followed increases in the rate of population growth, although the timing is not without dispute.

Population Induced Specialization in Agriculture

While population growth potentially augments the benefits of private property, potential efficiency gains do not automatically induce institutional change. In particular, rent seeking may lead to a ‘race’ such that private property is created before it actually increases efficiency (Lueck 1998). On the other hand, political costs may retard institutional change beyond the time that its benefits warrant. The advent of private property in Hawaii in 1848 was exceptional in two regards. First, the benefits of private property resulted from the increased profitability of sugar and pineapple production, even in the face of population decline. Second, the timing of private property accorded roughly with its efficiency benefits; the delaying effects of the political costs of change were offset by the expediency of governmental land sales.

A more profound institutional change that may be induced by population pressure and other sources of intensification is that of economic organization. The division of labour has fascinated economists since the time of Adam Smith, but was sidelined during the era of neoclassical economics. The theme of specialization has been resurrected, implicitly in endogenous growth theory and explicitly in the New Classical Economics (as in Yang 2003). In Yang’s model, population growth lowers the relative price of labour, thereby increasing the use and number of intermediate capital goods, which are produced with labour. This in turn increases production and the number of manufactured goods, and further bolsters the value of total output through learning-by-doing. In this model, agricultural growth is only indirectly stimulated, for example through the lower cost of manufactured fertilizer – a land- saving input.

Population growth can also facilitate specialization by lowering unit transaction costs. For example, the fixed costs of transport and communication infrastructure per capita may fall sufficiently to warrant additional infrastructure investment. Falling unit transaction costs, in turn, lower the friction that inhibits both horizontal and vertical specialization. In this case, learning-by-doing can directly bolster agricultural productivity.

A primary vehicle for increased specialization is hired labour. To see how population growth can induce hired labour, consider a hypothetical land-surplus economy wherein food is produced by family farms and where clearing costs are negligible. If we assume for the moment that output per hectare is a function of labour, farm size is efficiently determined where the marginal product of land is zero and the marginal product of labour is equal to the shadow price of household leisure. Once population growth brings lower quality, or sufficiently distant, land into production, intensification begins – lowering labour productivity. As the optimal land-to-labour ratio falls, the size of the average family farm declines. This process is efficiently halted, however, due to indivisibilities such as those associated with ploughs and draft animals. Eventually, farm size shrinks to a point where the economies of scale lost from further shrinkage are just offset by the transaction costs of hired labour. At this fundamental turning point, increases in labour per hectare induced by population growth are accommodated by hired labour instead of falling farm size. In this sense, the change in agricultural organization – known as the emergence of the rural proletariat – is not necessarily an indication of exploitation or inefficiency.

But hired labour is not a perfect substitute for family labour. Transaction costs are different, and, since hired labour is not necessarily tied to a particular farm, it can specialize in particular skills instead of adjusting to the attributes of that farm. In the common case where family labour has a higher shadow price of leisure, hired labour has a comparative advantage in arduous and well-defined tasks wherein transaction costs are manageable (for instance, because the results of the work are readily observable) and wherein speed and quality are enhanced by training and repetition. Family members have a comparative advantage in management-intensive tasks such as chemical applications that require knowledge of farm attributes and for which shirking is harder to control. The advent of hired labour stimulates horizontal specialization across tasks, as with men in the Philippines who specialize in transplanting rice and move from village to village to do so. The resultant learning- by-doing increases productivity – for example, in producing straighter rows of rice, which raise the productivity of workers through the use of rotary weeders. Vertical specialization also increases. For example, landowners may specialize in land improvements, such as irrigation, and employ tenants who specialize in management-intensive labour and who employ and monitor workers who specialize in arduous and more easily supervised tasks.

Further vertical and horizontal specialization is illustrated by the institution of piece-rate by teams. A team is hired to complete a task, such as transplanting, which is easily monitored by *ex post* inspection. In this sense, the task is equivalent to an intermediate good. The team may produce, for example, a stack of cane stalks that are of uniform length and ready for planting. Moreover, the team constitutes a separate firm. Its chief executive officer is the team manager, who contracts with the sugar grower and who bears the adverse reputational effects of any sub par performance. In this sense, the capacity for specialization in industry may be quantitatively greater than that of agriculture but not necessarily qualitatively different. Thus it is neither inevitable that population growth decreases or increases productivity in an agricultural economy.

The following stylized pattern of hired labour, based on Philippine rice farming in the 1960s to the 1980s, may serve to epitomize the evolution of specialization as labour intensification follows population growth. Once population density warrants clustered villages of farm families, the institution of exchange labour emerges for transplanting, harvesting, threshing, and often ploughing. Boserupian intensification increases the value of timeliness, and exchange labour allows these tasks to be completed in a day or less for one farm. The first widespread form of hired labour was for harvesting. Harvesters were paid a share of the harvest, typically one-sixth. This later evolved into the gama system, whereby a family or small group was assigned a portion of the farm to weed and later harvest, albeit for the same one- sixth share. This corresponded to a fall in wages relative to rents. In Java, Indonesia, where population pressure was even more intense, this same institution emerged – for the same one-sixth share – but the work requirement expanded even further, typically including transplanting.

When wage labour first appeared in Philippine rice farming, a given worker would typically perform a myriad of tasks over the cropping season. As intensification proceeded and the man-hours of hired labour increased, this undifferentiated wage-worker system was partially replaced by one involving specialized piece-rate workers who were paid according to their performance of a specific task. This evolved further into the piece-rate-by-team system described above. As per-hectare yields continued to increase, piece-rates were often converted back to wage contracts – due to the increased value of quality shirking – but task-by- task specialization was retained.

A common assertion in development economics is that large farms that rely primarily on hired labour are at a transaction-cost advantage relative to small, family farms. This view implicitly takes the distribution over farm size as exogenous, however. In the efficiency view sketched above, farm size is endogenous and responds to changes in population. Indeed, efficient farm size may actually increase as the increased incidence of hired labour warrants new contracting institutions that lower transaction costs. The transaction costs that remain are the necessary cost of retaining economies of scale and facilitating specialization. Whether productivity gains from specialization are enough to offset diminishing returns to more labour on a fixed amount of aggregate land cannot be determined a priori.

The view that share tenancy is inefficient is similarly incomplete. In the canonical view, share contracts are a pair-wise efficient institution for mitigating both the labour-shirking disadvantages of wage contracts and the risk-bearing disadvantages of rent contracts. Nonetheless, share tenancy is said to be socially inefficient because of the Marshallian labour shirking that remains under the common 50 per cent sharing. This view fails to explain how share tenancy fits into the evolution of agricultural organization in response to population pressure and other forces of intensification. Specialization is warranted by intensification and is facilitated by the evolution of contracts and other institutions. In particular, share tenancy facilitates vertical specialization between the landowner, the tenant, and the hired labour that the tenant supervises. It also facilitates the horizontal division of labour described above. On the other hand, share tenancy is primarily a type of family farm and may become less appropriate as agriculture becomes more capital-intensive. In any case, assessing the consequences of institutions without considering their causes, especially intensification, runs a risk of misplaced exogeneity.

A third example of questionable exogeneity concerns the view that the *modernization triad* – population pressure, technical change and commercialization – has inevitably immiserizing consequences. The case made against the new varieties of rice and wheat that emerged in the mid- to late 1960s is illustrative. Modern rice varieties are said to be most profitable on irrigated, highly productive land and for farmers facing relatively low shadow prices of credit and close connections with the money economy. These characteristics tend to favour wealthy landowners over small farm families. As the rich get richer, small farmers and tenants are allegedly disenfranchised, thus accelerating Ricardian forces of population and polarizing society into a class of landlords and the proletariat. Commercialization further augments proletariatization, breaking down safety-net customs such as gleaning rights for the poor, and setting the stage for violent conflict.

The Boserupian and induced innovation perspectives provide a compelling counterweight to the neo-Marxian view. Technical change induced by population growth is primarily land-saving and offsets downward wage pressure, whereas Marxian technical change is strongly labour-saving and exacerbates the downward effect of population. Like induced technical change, induced institutional change in the form of 'commercialization' has a positive effect on wages. The efficient emergence of landless workers helps to avoid the immiserizing effects that would occur from a growing population being accommodated by shrinking farm sizes. This class division in turn creates both a supply and a demand for hired labour. As labour markets emerge, new institutions such as piece-rate contracts and work teams with team leaders emerge to lower contracting costs, thereby lowering the transaction cost wedge between effective wage paid, including costs of recruitment, training and supervision, and effective wage received, net of the costs of search, required tools, and the journey to work. As the unit-transaction-cost wedge shrinks, workers move up their supply curves and employers down their demand curves for labour, resulting in more hired labour and increased net wages. From this perspective, induced innovation at least partially offsets the downward pressure that population pressure puts on wages.

These efficiency patterns are by no means inevitable, but serve to counter the view that the modernization triad is inevitably impoverishing. The efficiency view also provides a theoretical starting point for explaining agricultural growth or the lack thereof. Rent-seeking and policy distortions may induce arbitrary and inefficient patterns of ownership and farm size, thereby inhibiting the efficiency forces described. A challenge for economic historians and agricultural development theorists is to explain the political-economy forces that have facilitated induced innovation in some cases and inhibited it in others.

The positive Boserupian forces of induced innovation and specialization move in the opposite direction of the classical Malthusian effects. To summarize the above, even a small family farm can have four levels of vertical specialization – landowner, share-tenant farm manager, work team leader, and worker – as well as horizontal specialization across the array of farm tasks. The advent of each new form of specialization can be modelled along the lines of Fig. 1. Because of the non-convexity associated with the fixed cost of each advance in organizational complexity, population-induced specialization gives rise to increased labour productivity, but only over a limited range of additional labour. In the absence of other effects and changes, we would expect to see the marginal and average products of labour initially rising after each increase in specialization; then, as labour per hectare increases further, to a decline until the next innovation is made. Adding learning-by-doing to the picture increases the chances of sustained productivity gains. Nonetheless, the theory cannot tell us whether the positive forces will outweigh the negative Malthusian forces in the long run.

A Historical Perspective

The history of agricultural growth is informative. As documented by Evans (1998), the long-run rate of agricultural growth closely matched that of population until 1825, when world population reached one billion people. The corresponding increase in food production was almost entirely sourced in an increase in cultivated area, that is, it was *extensive* in nature. In contrast, since world population reached five billion late in the 20th century, the increase in food production has been almost entirely driven by increased productivity. During the intervening period, when world population increased by four billion, growth in food production was increasingly intensive in nature (due to increased inputs) with increased productivity becoming more important as the period progressed. That is, as intensification led to diminishing returns, increased productivity became increasingly important.

This broad-brush generalization about the nature of agricultural growth is consistent with the induced innovation perspective. As population growth increases land scarcity, the Ricardian gradient, which depicts the proportion of agricultural growth due to intensification, is monotonically rising. Intensification increases the relative scarcity of land further, relative to labour and capital, thus stimulating induced productivity increases, both from technical and institutional progress. Ironically, food supply has grown 'geometrically' since 1938 (averaging 2.2 per cent per year) and population has grown nearly 'arithmetically' since 1959 (with one billion being added to world population roughly every 13.3 years). Technological and institutional change has seemingly inverted Malthusian theory.

This does not imply that all technological change is demand-induced. Even the theory of induced innovation admits supply-side innovations. For example, knowledge capital produced in the defence industry may lead to better communications technology. Irrigation systems in ancient Mesopotamia and Egypt were presumably not induced by increasing land scarcity but because someone figured out how to produce more with less. Economic history in the United States suggests that demand was partly induced by labour scarcity, but, once certain types of farm equipment had been invented, they were adapted even in areas where land prices were increasing faster than labour prices. Kremer (1993) even suggests that until the late 18th century the Malthusian argument was so predominant that population could be viewed as a proxy for technological change.

On the other hand, the agricultural and industrial 'revolutions' are now viewed less as bursts in productivity spurred by invention and more as induced technical change. For example, the four-field system, whereby wheat, barley, turnips and clover were grown in separate fields and rotated the following year, was once viewed as an essential part of the English agricultural revolution during the 18th century. But the system was developed in land-scarce Flanders two centuries before and popularized in England only once it was warranted by sufficient population-induced land scarcity.

Even the mechanism of induced technical change is not entirely governed by factor prices, however. For example, the replacement of the fallow period in the medieval 'three field' rotation by beans or another leguminous crop appears to have been indirectly induced by the population decline in 14th-century western Europe. Higher wages and farm incomes, resulting from the lower population and decreased land scarcity, increased the demand for meat. Complemented with the Flemish demand for wool, this incentivized farmers to increase sheep production, and they responded by both converting some lands to pasture and growing legumes in place of fallow on much of the remaining lands.

The extent to which technical change in English agriculture was induced has been the subject of intense historical debate. Historians reporting that agricultural productivity increased rapidly, say in the late 18th and early 19th centuries, tend to see an agricultural revolution stimulated by exogenous technical change. Economic historians who estimate productivity increases to be quite gradual view changes in rotation and other innovations as induced. As suggested by the discussion of Fig. 1, induced changes do not by themselves reverse the price and income trends that induced them in the first place and therefore tend not to be associated with dramatic increases in productivity.

Sustainable Development

Resource depletion adds another negative dimension to the never ending debate between the development optimists and pessimists. Even before *sustainable development* became fashionable, neo-Malthusians argued that unbridled population growth in poor countries and economic growth in rich countries must inevitably cause severe pressure on the earth’s limited resources, resulting in burgeoning poverty and international conflict. The only solution was said to be the *steady state economy* with constant population, capital stock, and output.

After the Brundtland Commission’s 1987 report, resource depletion was broadened to include pollution and other environmental threats. Environmental degradation, including increasing water scarcity, soil erosion, deforestation, desertification, salinization, and global warming, as well as diminishing energy and marine resources, was viewed as exacerbating the Malthusian vicious circle. Accordingly, the Brundtland Commission called for a simultaneous assault on population growth, poverty and environmental degradation, thus giving rise to the modern movement for *sustainable development.* Economists have had limited success in modelling sustainable development, however. One notable review and synthesis (Arrow et al. 2004) was unable to settle on positive principles of sustainability and settled on the negative *sustainability criterion* – an injunction not to deplete the value of natural capital more than the additional value of produced capital.

Even if we abstract from technical change, expanding models of economic growth to include environmental degradation does not produce a necessarily dismal outlook, however. If we represent concern for future generations by *intergenerational neutrality* and assume that population grows exponentially at a constant rate, optimal per capita consumption grows to its *golden rule* level, under plausible assumptions about substitutability, both between renewable and non-renewable resources and between natural and produced capital. Adding technical change provides even rosier possibilities (Weitzman 1997). Whether these possibilities are realized depends largely on the effectiveness of private and public governance structures in facilitating specialization and exchange while guarding against unproductive rent seeking (Greif 2006).

The Co-evolution of Specialization and Governance

The economic history of Hawaii provides a relatively recent, pre-industrial example of how specialization and governance in agriculture co-evolve with changes in population. During the ‘colonization’ period AD 300–600, population growth, including further migration of Polynesian peoples, was slow. Agricultural expansion was extensive. The population began to increase more rapidly towards the latter part of the ‘development’ period (600–1100), and agriculture began to intensify with the advent of irrigation. There was little if any division of labour among the commoners. During the expansion period (1100–1650) population accelerated and intensification greatly increased with a decreased fallow period, a major expansion in irrigation and with the development of fishponds. Horizontal specialization among workers became commonplace, with fishing more of a distinct occupation. Evolving from a system of somewhat separate extended families units, social and production relations became increasingly stratified, eventually with a distinct hierarchy from local chief upwards to governor *(ali’i)* of the watershed to district head (see Kirch 1985).

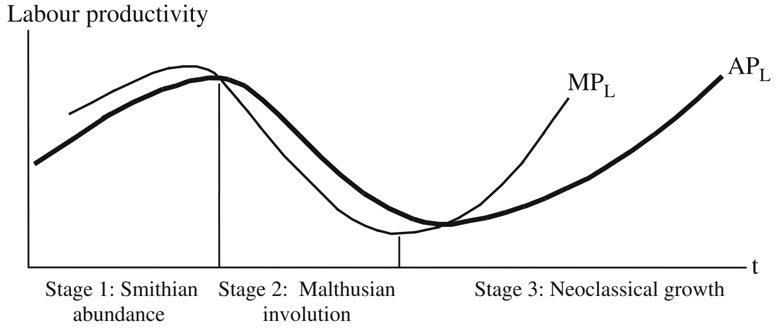
This stylized history is suggestive of a *governmental Kuznets curve.* During the extensive (pioneer) stage of development, family or extended family units are largely autonomous and decision-making is decentralized accordingly. During the intensive development stage, decision-making and governance are centralized at a higher, albeit intermediate level (for example, communal governance of the commons). As intensification and specialization continue, efficiency favours a further centralization of governance, at least for the minimal functions of defence and the justice system, but a decentralization of decision-making as facilitated by private property. This last stage occurred in Hawaii after Western contact in 1778. New trade opportunities raised the value of irrigation and other investments in plantation agriculture, initially for sugar and later pineapple. Private property provided the assurance that planters needed to commit to these investments and also facilitated specialization between districts that was warranted by international trading opportunities. Graphing this historical progression of increasing governmental centralization on the horizontal axis, and rising and then falling centralization of decision-making on the vertical axis, completes the governmental Kuznets curve. Viewing government intervention in these two dimensions provides a useful antidote to the misleading question of 'how much government' that sometimes arises in policy circles.

Smith to Malthus to Solow

A largely unexplored area of enquiry involves combining the theory of endogenous population growth with the theory of sustainable growth outlined above. Perhaps the simplest model of endogenous growth can be found in two-sector growth models of economic development wherein the birth rate is exogenous and the death rate declines to minimum as per capita income increases. The birth rate may also be made endogenous following the Chicago School's new household economics. The increased opportunity cost of child care is one pervasive cause of the decline in fertility with economic development. Moreover, as the capital intensity of the economy increases, the returns to human capital are raised, thus creating incentives for families (individually or collectively) to invest in human capital, a partial substitute for increased fertility.

Malthus's emphasis on the supply of food determining population and Boserup's focus on exogenous population growth increasing the demand for land and inducing supply side changes in agricultural production are clearly complementary. Focusing on one or the other is a device for dealing with the shortcomings of human imagination and the fact that models with both forces are indeterminate without further, possibly arbitrary, restrictions added to the model. Indeed, due to the endogeneity of population, enquiring into the impact of population levels involves something of a category mistake. In light of this, the World Bank statement (1984; see also Kelley 1988) that population growth in excess of two per cent per annum tends to have a negative impact on per capita income warrants reinterpretation. A more accurate statement would be that population growth in excess of two per cent tends to be associated with negative growth in per capita income after partially controlling for (imperfectly measured) positive effects. In particular, where high population growth occurs in the face of policy failures that cause an anti-labour bias, population growth tends to exacerbate the Brundtland vicious circle described above.

More generally, the effects of population growth on agricultural and economic development may be different depending on the population density and the stage of economic development, as illustrated in Fig. 2. For the early American frontier and for parts of Africa today, physiological population density may be sufficiently sparse for Smithian economies of specialization and Boserupian economies in infrastructure to afford increasing labour productivity, as shown by the rising segment of the average product of labour curve. There is no labour market, at least in the sense of a competitive spot market, in such economies because paying labour its marginal product would more than exhaust total output. When the extensive land frontier



Population and Agricultural Growth, Fig. 2 Stages of economic development the marginal product of labour begins to rise, causing wage rates to rise and pulling up average labour productivity soon thereafter. Accumulation of produced capital and the relative increase of the industrial sector generate the transition to modern economic growth

nears economic exhaustion, population density becomes high, and the economy is still dominated by agriculture (as on the Indonesian island of Java in the 1960s and early 1970s), real wages fall, along with the average product of labour. Once the 'structural transformation' takes place, such that the growth rate of the agricultural labour force (if any) is but a small fraction of that of the industrial labour force.

These stages are not inevitable forces of history. Some economies may be able to bypass the Malthusian stage altogether. For example, economic policies in Taiwan during the 1950s and 1960s encouraged labour intensity in agriculture. This and the investments in physical infrastructure, a gradual transition to processing and high value-added agricultural production and an efficient system of marketing cooperatives kept the demand for labour and wages rising. Hong Kong and Singapore were able to skip the Malthusian stage by early industrialization that relied on trade instead of the Johnston–Mellor linkages whereby agricultural development increases incomes (thus stimulating demand for industrial products), mobilizes savings for industrial investments, and provides a market for manufactured farm inputs (Johnston 1970). Korea was similarly able to bypass an extended Malthusian stage by allowing investment coordination through *chaebols* (business groups)and focusing on manufactured exports. In contrast, the negative force of policy failures can extend Malthusian involution and even prevent the transition to modern economic growth. Finally, because of policy failures and exogenous shocks, history may record more than two turning points. For example, after going through a Malthusian period during the 'long 16th century', wages in England rose between approximately 1640 and 1740, but then fell again before entering a 'Solovian' period of increase starting slightly after the advent of the 19th century and accelerating after the American Civil War.

Nonetheless, we may meaningfully enquire into the mechanics of the two turning points shown, after abstracting from policy failures and exogenous shocks. While the first turning point has clear Ricardian underpinnings, the second has generated substantial controversy. How does an economy go from 'Malthus to Solow?' Forward linkages from agriculture are important in explaining the relative growth of industry, but they do not, in and of themselves, explain the rapid and sustained growth in labour productivity during modern economic growth.

Note first that there is an implicit Kuznets curve corresponding to Fig. 2. During the Malthusian period, wages fall and Ricardian rents increase, worsening income distribution. Even as industrialization begins to pull up wages, income distribution may continue to worsen for some time as the total returns to capital increase faster than the wage bill. Eventually, as the returns to human capital induce the substitution of 'quality for quantity' in fertility decisions, widely distributed human capital accumulates and even produced capital becomes less concentrated. These forces cause a more equal income distribution in the model.

Were it only for Ricardian landlords accumulating an agricultural surplus and financing industrialization and the production of goods for a landed aristocracy, industrialization would have not have been as robust as that witnessed in modern economic growth. Indeed, increasing wages stifle the labour-intensive production that characterizes the early stages of industrialization, decrease the agricultural surplus, and detract from the rental incomes of capitalists and landlords that finance capital formation. What saves the day are the non-convexities inherent in industrialization.

While there are numerous possibilities for specialization and other non-convexities in agriculture, these are still few in comparison with those in industry. In industry, there is more horizontal specialization through proliferation in the number of products and more vertical specialization through multiple stages of intermediate production. In agriculture, the number of products is more limited, and vertical specialization without industry tends to be limited to separation of management and labour. With industry, agriculture can take advantage of land-saving intermediates such as fertilizer and tractors. Thus it is plausible that technological and institutional changes in agriculture have not been frequent enough to overcome the inexorable Malthusian force of increased food affording greater population growth.

In contrast, once industry becomes a major part of the economy, non-convexities may be sufficiently compact in the course of development to dominate the negative force of lower death rates. The resultant increase in per capita income in turn invokes a positive feedback mechanism whereby Engel effects increase the demand for manufactures, thus increasing capital formation and the returns to human capital, thereby contributing to the decline in the demand for child numbers described above. Greater product specialization and falling unit transport costs afford a further inducement to international trade, an additional positive feedback mechanism. This theory supports the revisionist interpretation that the agricultural and industrial 'revolutions' were misreadings of a gradual process of economic change (see Clark 2007).

The role of industrial development in sustaining increased wages and per capita incomes does not imply that the appropriate development policy requires pushing industrial development while 'squeezing' or neglecting the agricultural sector. Indeed, for countries with a preponderance of the labour force in agriculture, economic development can be sustained only by 'pushing' on the agricultural sector with R&D, infrastructure, and non-confiscatory prices (Pingali 2006). It does mean, however, that stimulating the agricultural sector alone – that is, relying on automatic linkages from the agricultural to the industrial sector – is not sufficient for sustained economic development. External economies of labour-market pooling, human capital, technological spillovers and other network externalities imply that there are aspects of investment coordination that are not internalized by spot markets. This leaves an important role for government in facilitating the requisite economic cooperation.

See Also

▶ Agriculture and Economic Development
▶ Institutional Economics

Bibliography

Ahmad, S. 1966. On the theory of induced invention. *Economic Journal* 76: 344–357.

Arrow, K.J., P. Dasgupta, L.H. Goulder, G.C. Daily, P. Ehrlich, G. Heal, S. Levin, K.-G. Maler, S.H. Schneider, D. Starrett, and B. Walker. 2004. Are we consuming too much? *Journal of Economic Perspectives* 18(3): 147–172.

Barrett, S. 1991. Optimal soil conservation and reform of agricultural pricing policies. *Journal of Development Economics* 36: 167–187.

Barrett, C.B., T. Reardon, and P. Webb. 2001. Non-farm income diversification and household livelihood strategies in rural Africa: Concepts, dynamics and policy implications. *Food Policy* 26: 315–331.

Binswanger, H.P., and V.W. Ruttan. 1978. *Induced innovation: Technology, institutions, and development*. Baltimore: Johns Hopkins University Press.

Boserup, E. 1965. *The conditions of agricultural growth*. London: Allen & Unwin.

Boserup, E. 1981. *Population and technological change: A study of long-term trends*. Chicago: University of Chicago Press.

Boserup, E. 1987. Agricultural growth and population change. In *The new palgrave: A dictionary of economics*, vol. 1, ed. J. Eatwell, M. Milgate, and P. Newman. London: Macmillan.

Brundtland Commission. 1987. *Our common future*. Oxford: Oxford University Press.

Clark, G. 2007. *A farewell to alms: A brief economic history of the world*. Princeton: Princeton University Press.

Evans, L.T. 1998. *Feeding the ten billion: Plants and population growth*. Cambridge: Cambridge University Press.

Evenson, R.E. 1982. *The green revolution in north Indian agriculture: Ex post assessment*. Mimeo: Yale University.

Evenson, R.E. 2004. Food and population: D. Gale Johnson and the green revolution. *Economic Development and Cultural Change* 52: 543–570.

Greif, A. 2006. *Institutions and the path to economic modernity: Lessons from medieval trade*. Cambridge: Cambridge University Press.

Hayami, Y., and V.W. Ruttan. 1985. *Agricultural development: An international perspective*. Baltimore: Johns Hopkins University Press.

Johnston, B.F. 1970. Agriculture and structural transformation in developing countries: A survey of research. *Journal of Economic Literature* 8: 369–404.

Kelley, A.C. 1988. Economic consequences of population change in the third world. *Journal of Economic Literature* 26: 1685–1728.

Kirch, P.V. 1985. *Feathered gods and fishhooks: An introduction to Hawaiian archaeology and prehistory.* Honolulu: University of Hawaii Press.

Kremer, M. 1993. Population growth and technological change: One million BC to 1990. *Quarterly Journal of Economics* 108: 681–716.

Krautkraemer, J. 1994. Population growth, soil fertility, and agricultural intensification. *Journal of Development Economics* 44: 403–428.

Lucas, R.E. 2002. *Lectures on economic growth*. Cambridge, MA: Harvard University Press.

Lueck, D. 1998. First possession. In *The new palgrave dictionary of economics and the law*, ed. P. Newman. New York: Macmillan.

Malthus, T.R. 1798. *An essay on the principle of population*. Oxford: Oxford University Press, 1999.

North, D.C., and R.P. Thomas. 1973. *The rise of the western world: A new economic history.* New York: Cambridge University Press.

Pender, J. 2001. Rural population growth, agricultural change and natural resource management in developing countries: A review of hypotheses and some evidence from Honduras. In *Population matters: Demographic change, poverty and economic growth in developing countries*, ed. N. Birdsall, S. Sinding, and A. Kelley. Oxford: Oxford University Press.

Pingali, P. 2006. Agricultural growth and economic development: A view through the globalization lens. Presidential Address, International Association of Agricultural Economists 26th Conference, Queensland.

Ricardo, D. 1817. *On the principles of political economy and taxation*. London: John Murray.

Roumasset, J. 2004. Rural institutions, agricultural development, and pro-poor economic growth. *Asian Journal of Agriculture and Development* 1: 61–82.

Ruttan, V.W., and Y. Hayami. 1973. Technology transfer and agricultural development. *Technology and Culture* 14: 119–151.

Weitzman, M.L. 1997. Sustainability and technological progress. *Scandinavian Journal of Economics* 99: 1–13.

World Bank. 1984. *World bank development report: Population change and development*. Oxford: Oxford University Press.

Yang, X. 2003. *Economic development and the division of labor.* New York: Blackwell.

Population Dynamics

Ronald D. Lee

Abstract

Population dynamics are the patterns of change over time in populations. Populations fluctuate in response to fluctuating external forces, or because of the internal structure of the process of demographic renewal. Damped cycles one generation long may result from the interaction of random perturbation and the age distribution of reproduction. So-called Easterlin cycles two generations long, either damped or self-exciting, may arise from the lag between birth and labour force entry when fertility responds sensitively to labour market conditions. Longer-term dynamics arise from the interactions of population growth, capital, endogenous technology, and income.

Keywords

Easterlin cycles; Easterlin hypothesis; Fertility; Kondratieff cycles; Kuznets swings; Malthus, T.; Malthus's theory of population; Net maternity function; Net reproduction rate; Population and economic growth; Population density; Population dynamics; Solow, R.; Stable population theory; Technical progress

JEL Classifications

J10

Population dynamics are the patterns of change over time in populations, ranging from fluctuations to long-term trends, and the underlying principles that govern these changes.

Population Fluctuations

All human populations exhibit fluctuations in their vital rates and consequent irregularities in their age distributions to a greater or lesser degree. Analyses of such fluctuations are of interest for many reasons – for historical understanding, as a basis for forecasting, for a deeper understanding of underlying social processes – but perhaps most intriguing is the possibility that they may afford some insight into more fundamental aspects of population dynamics and may illuminate the very process of demographic renewal. More specifically, we may be able to learn from the occurrence or absence of longer cycles whether a population is subject to negative feedback of a Malthusian sort and perhaps to place bounds on its sensitivity if it occurs. To Malthus (1798) it seemed obvious that populations would perpetually oscillate about equilibrium. This notion is taken seriously as an interpretation of the long swings in the fertility of many contemporary developed countries, as we discuss in more detail below.

Fluctuations may come about in three ways (or through combinations of these ways). First, they may simply be imposed on a series of births or deaths by fluctuations in some driving force such as prices or the weather. In this case, both the amplitude and the period of the fluctuation depend entirely on the driving series. Second, damped fluctuations may be created by the internal structure of a demographic process, as it responds to random and non-cyclic external shocks; in this case the cycles will die out if the external disturbance stops. The period of such cycles depends entirely on the nature of the renewal process, not on the driving force; however, the amplitude of the cycles depends on the amplitude (variance) of the disturbing force.

The third possibility is that limit cycles occur. Like the aforementioned cycles, these are generated by the internal structure of the reproductive process, but unlike them they are self-sustaining or ‘self-exciting’ and would continue indefinitely even in the absence of outside shocks. In this case, both the amplitude and the period depend only on the reproductive process. When a dynamic equilibrium is unstable, such that trajectories tend to explode away from the equilibrium path, then one of three things may happen: explosive fluctuations may lead to extinction; the non-repeating fluctuations of chaos cycles may occur; or the system may settle down to a limiting pattern of cycles, called limit cycles. There are many examples of animal populations exhibiting such behaviour. In human demography, it is a matter of controversy whether such cycles have ever actually occurred, but if they have it is presumably through the kind of mechanism proposed by Easterlin (1968), a sort of Malthusian cycle about equilibrium.

Imposed Cycles

There are well known non-seasonal cycles in fertility and mortality at or below the annual frequency (obstetricians avoid deliveries on Sundays; people have lower mortality just before elections, compensated for by increased mortality thereafter). Seasonality is strong in fertility, mortality, nuptiality and migration, particularly in traditional agricultural societies and in those less insulated by their dwellings from the variations of climate (in the extreme case of Bangladesh in the 1970s, for example, the seasonal peak in fertility was two to three times the seasonal trough). In the case of mortality, nuptiality and migration the causes of seasonal variation are fairly well understood to be rooted in identifiable biological, institutional and economic influences. In the case of fertility, the causes of seasonality are much less well understood.

There are also somewhat longer fluctuations in vital rates, in the range of 2–15 years. These have been quite thoroughly studied and found to be associated with business cycle indicators in the developed world and with the harvest cycle in preindustrial conditions. Lower agricultural prices and less unemployment are associated with higher fertility and nuptiality and with lower mortality, with lag patterns of response indicating that much of the variation is confined to changes in the timing of events. Fluctuations in temperature also are important, with colder winters and hotter summers raising mortality and reducing fertility, with an appropriate lag. In the case of mortality, exogenous epidemiological variation historically played a larger role (Wrigley and Schofield 1981). These relationships have continued to hold at least until a few decades ago in the developed countries and are still evident in the Third World countries where they have been investigated.

Much longer fluctuations in population variables are also visible in the historical record. Kuznets cycles, of 15–25 years, include a procyclical response of migration, both internal and international. Some of the birth-rate series of 19th century Europe show signs of the Kondratieff cycle. But most striking are the waves lasting two or three centuries in the demography of Europe and of China, from at least the 12th century up to the 18th (Wrigley and Schofield 1981). These are evident in population growth rates and in mortality; their existence in fertility is problematic. The cause of these very long waves is not clear, although a case can be made for the influence of climatic variation and for the effects of intercontinental exchange of diseases through conquest or trade. Whatever their cause, such demographic fluctuations played a critical role in economic history, driving rents, wages and other relative prices, and possibly inflation. It is possible that such fluctuations were generated internally by the economic demographic system as Malthusian fluctuations about equilibrium; in the present state of knowledge, however, it appears more likely that the cycles were imposed.

Cycles Arising from the Internal Age and Temporal Structure of Reproduction

A characteristic pattern of delay between an event and its recurrence can act as a filter which creates quasi-cyclic behaviour in the series of events when the timing is subject to continual random perturbation. In this way, the typical spacing of a mother's births two to three years apart tends to generate cycles of this length, as was first pointed out by Yule (1906). Such cycles are visually discernible in many birth and fertility series and show up in the empirical power spectra.

More importantly, the typical delay between a woman's own birth and the time she herself gives birth to female children leads to cycles of 25–35 years, or the approximate length of a generation, when fertility is randomly perturbed (see Coale 1972; Lee 1974). This may be shown as follows. Let B(t) be the number of births in year t, and let $\varphi(a)$ be the expected number of births to each of these births at age a, net of mortality ($\varphi(a)$ is known as the 'net maternity function'). $\varphi(a)$ typically rises from zero at an age around 15 years to a peak in the twenties and declines again to zero at around age 45; its mean, μ, is the mean age at child-bearing and falls between 25 and 35 years depending on the population. The renewal process is written:

$$B(t) = \sum \varphi(a) B(t-a) \qquad (1)$$

where the sum is taken over the reproductive years. Such a process will settle down to a stable exponential growth path if the characteristic roots of φ lie within the unit circle. But as the B series converges to this growth path from an irregular past, it will fluctuate, and the fluctuations can be characterized by further examination of the characteristic roots of φ. There will generally be one real root, describing the steady state growth rate, and the others will come in pairs of complex conjugates. The pair with the largest modulus is the only one of substantive interest; it will describe a damped oscillation with length roughly equal to the mean age of child-bearing, μ. Any initially distorted age distribution, if subsequently subjected to fixed vital rates described by ϕ, will generate a birth sequence which moves in waves one generation long as it converges towards exponential growth.

The argument can easily be generalized to cover the case of a population whose net maternity function is subject to constant stochastic disturbance of any autocovariance structure; the age structure of reproduction, described by the mean values of φ, will amplify variation in the neighbourhood of frequencies corresponding to cycle length μ, leave them unchanged at higher frequencies, and attenuate them in the neighbourhood of cycle length 2μ. Thus, a population in a random environment will tend to exhibit cycles one generation long or to superimpose these on whatever pattern of variation is forced on it by the environment. The birth series of many pre-industrial populations, particularly at the parish level, indeed do reveal such waves; whether the mechanism described above suffices to account for them has not yet been established empirically.

Some scholars have seen a major economic influence in such population waves, but this view now appears exaggerated; waves generated in this way are generally quite mild; they have low amplitude, and they damp fairly rapidly following an identifiable disturbance.

Cycles Arising from Economic-Demographic Interaction

Interest in dynamic economic–demographic models of population renewal, stressing fluctuations arising from age distributions, was prompted by the long 'cycle' in US fertility, with a trough in the 1930s, a peak in the late 1950s, and a trough in the 1970s. A number of scholars, most notably Easterlin, suggested around 1960 that the fertility fluctuations might reflect the economic conditions faced by young labour market entrants, conditions which in turn were worse for large cohorts and better for smaller ones. This insight led them to forecast correctly the sharp decline in fertility occurring in the 1960s, as larger cohorts aged into the labour market. Easterlin (1968) developed a detailed theory, buttressed by extensive empirical investigation, leading to a tentative prediction of self-generating demographic cycles two generations long, as small birth cohorts had high fertility and gave birth to large cohorts, who in turn reared small cohorts, and so on. Such cycles are known as 'Easterlin cycles'. A considerable empirical literature has since appeared on the subject, lending considerable support at the aggregate time series level in the United States and some other countries, but very little at the micro level.

I now briefly review the theoretical literature on economic–demographic cycles. The account of the renewal process given above implicitly assumed that net maternity at time t, $\varphi(a, t)$, was independent of the population age distribution at time t, or equivalently of the preceding series of births. But it is entirely possible that this is not so. Suppose, for example, that a Malthusian model is appropriate, such that fluctuations in labour supply lead to inverse fluctuations in wages, and that fertility depends positively on the wage level. This leads to different dynamic possibilities and a modified renewal equation.

Suppose that the net maternity function, $\varphi(a)$, depends on some set of economic variables, let us say wages for concreteness. Suppose that these in turn depend on some set of economic variables, Z, which are independent of age distribution, as well as on the current population age distribution, which thus in conjunction with Z determines wages. If mortality is constant and the population closed to migration, as we here assume, then the current age distribution is completely determined by past births. We can then write:

$$B(t) = \sum \varphi^{*}[B(t), Z(t)]\, B(t-a), \qquad (2)$$

where $B(t)$ denotes the vector of past births; this replaces the purely demographic renewal eq. (1) introduced above (Lee 1974).

The renewal process will have an exponential equilibrium growth path, $B^{*}(t) = B\, exp(nt)$, which satisfies (2) for all t. For simplicity, suppose that Z is such that $n = 0$, so that the equilibrium path is stationary. It is helpful to consider the process of proportional deviations about this equilibrium path, denoted $b(t)$. Let $\varphi(a)$ be the value of $\varphi^{*}[\ ]$ evaluated at equilibrium. In this case, the sum of $\varphi(a)$ over all a, known as the net reproduction rate or NRR, is unity when evaluated at the equilibrium age distribution. Let $\Gamma(a)$ be the elasticity of the NRR with respect to the size of age group a, or equivalently with respect to births a years previously, $B\ (t\ -\ a)$; these elasticities are readily derived from the original function φ. Then the renewal process for fluctuations about the equilibrium growth path of births is simply:

$$b(t) = \sum [\varphi(a) + \Gamma(a)]b(t-a). \qquad (3)$$

The smaller the effect of the current age distribution on fertility (Γ), the more the population renewal process resembles the purely demographic version of (1). In any event, exactly the same procedures can be used to study the dynamic behaviour of birth fluctuations in this model as were used previously.

The first step is to check the characteristic roots to assess stability. If the oscillations of the process tend to explode away from the equilibrium growth path, then a different kind of analysis, discussed below, is called for. If the roots indicate that oscillations are damped, then the analysis of dynamic behaviour in the neighbourhood of equilibrium will be informative.

We can now consider specifications of the model which have been proposed in the literature. The first is the simplest Malthusian model, in which all age groups in the labour force are assumed to be perfect substitutes in production, and fertility at each age is assumed to be negatively related to the size of the potential labour force, through an hypothesized effect on wages. In this case, $\Gamma(\alpha) = \beta\, k(a)$, where β is independent of age, and expresses the sensitivity of response (elasticity of the net reproduction rate with respect to labour force size at equilibrium), while the $k(a)$ depend only on mortality conditions and equilibrium age specific labour supply and are therefore easily calculated from data at hand. Depending on values of β, this model will generate cycles ranging from one generation (as in the purely demographic model) to a century and a half or more. For $\beta = 7.5$, which is the empirical estimate from US data, 1917–1973, a cycle corresponding to the observed time path of births may be produced (Lee 1974; Wachter 1991).

Another model which is often used makes the fertility of a birth cohort depend only on the size of the cohort and makes it independent of all other age group sizes. The simplest form of this specification leads to:

$$b(t) = (1-\alpha) \sum \varphi(a)\, b(t-a), \qquad (4)$$

where α is the elasticity of each age's fertility with respect to cohort size. For α less than 1, there is a generation-long cycle; for α greater than 1 but less than 2, there is a damped two-generation cycle, and for α greater than 2 an explosive two-generation cycle occurs.

Specifications reflecting other degrees of substitutability of age groups of labour could of course be tried. Easterlin typically has used a ratio of younger to older workers to drive fertility (this could be derived from a CES model with two age groups of labour as separate factors, for example). The general expression can be used to explore dynamics under a wider variety of specifications. For example, the burden of supporting the elderly retired population might lead to a reduction in fertility; this would be expressed as a suitable negative $\Gamma(\alpha)$ for $a =$ 65 and over. If couples were led to desire larger families when they observed other couples' children, then $\Gamma(\alpha)$ would be positive for ages zero to ten.

When the cyclic behaviour near equilibrium is found to be explosive, then we need to consider behaviour further from equilibrium, at which point nonlinearities become important (unless, of course, the behaviour is truly linear, in which case population extinction results). Dynamic behaviour can be 'chaotic', an endless series of non-repeating fluctuations; for many models, however, limit cycles will occur, with amplitude and period determined not by the pattern of disturbances but rather by the functional relations themselves. Such cycles are observed in animal populations in laboratories and occasionally in the wild; in human populations their occurrence is conjectural: Samuelson (1976) considered a particular three-age group model leading to limit cycles.

Long-Term Population Trends and Economic Growth

Longer-term trends in population have also been viewed in the context of processes related to economic growth. Solow (1956) studied the behaviour of a population whose growth varied first positively and then negatively with respect to per capita income. Combining this study with his neoclassical growth model, he showed there was a stable low-level equilibrium at which per capita income was constant and population grew at the rate of technological progress, but also a second equilibrium at a high per capita income which was unstable. If the capital–labour ratio could be raised slightly above this equilibrium level, then per capita income would rise without limit while the population growth rate fell lower and lower.

In Solow's approach, as in Malthus's, technological progress was taken as exogenous. Boserup (1981) and others have suggested that larger denser populations would be more likely to experience technological progress in the long run, for reasons related to both the supply of innovations and the demand for them. She suggested that, combined with a Malthusian endogenous response of population growth to economic progress, an upward spiral of population growth and technological progress might occur, with positive feedback. A number of scholars have developed formal models of this process (Lee 1986; Kremer 1993), in a literature that overlaps slightly with the endogenous growth literature (Jones 2003).

See Also

- ▶ Easterlin Hypothesis
- ▶ Kondratieff Cycles
- ▶ Kuznets Swings
- ▶ Stable Population Theory

Bibliography

Boserup, E. 1981. *Population and technological change*. Chicago: University of Chicago Press.

Coale, A. 1972. *The growth and structure of human populations: A mathematical investigation*. Princeton: Princeton University Press.

Easterlin, R. 1968. *Population, labor force, and long swings in economic growth*. New York: National Bureau for Economic Research.

Jones, C. 2003. Population and ideas: A theory of endogenous growth. In *Knowledge, information, and expectations in modern macroeconomics: In honor of Edmund S. Phelps*, ed. P. Aghion et al. Princeton: Princeton University Press.

Kremer, M. 1993. Population growth and technological change: 1,000,000 B.C. to 1990. *Quarterly Journal of Economics* 108: 681–716.

Lee, R. 1974. The formal dynamics of controlled populations and the echo, boom and the bust. *Demography* 11: 563–585.

Lee, R. 1986. Malthus and Boserup: A dynamic synthesis. In *The state of population theory*, ed. D. Coleman and R. Schofield. Oxford: Basil Blackwell.

Lee, R. 1997. Population dynamics: Equilibrium, disequilibrium, and consequences of fluctuations. *Handbook of population and family economics*, v. 1B, ed. M. Rosenzweig and O. Stark. Amsterdam: North-Holland.

Malthus, T. 1798. In *An essay on the principle of population*, ed. A. Flew. Harmondsworth: Penguin. , 1970.

Samuelson, P. 1976. An economist's non-linear model of self-generated fertility waves. *Population Studies* 30: 243–247.

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

Wachter, K. 1991. Elusive cycles: Are there dynamically possible Lee–Easterlin models for US births? *Population Studies* 45: 109–135.

Wrigley, E., and R. Schofield. 1981. *The population history of England 1541–1871: A reconstruction*. Cambridge, MA: Harvard University Press.

Yule, G. 1906. Changes in the marriage and birth rates in England and Wales during the past half century. *Journal of the Royal Statistical Society* 69 (1): 18–132.

Population Health, Economic Implications of

David Canning and David E. Bloom

Abstract

Population health is not only a consequence but also a cause of a high level of income. Healthier people are more productive in work. Healthy children have better school attendance and cognitive development, while longer prospective working lifespans encourage investments in education. Longer lifespans can also increase saving and wealth accumulation as an extended retirement becomes more likely. The beneficial effects of population health can be seen both at the individual and macroeconomic levels, while the continuing high burden of disease in sub-Saharan Africa poses a substantial challenge to its economic development.

Keywords

Age structure; Black Death; Cognitive ability; Demographic transition; Dependency; Economic growth; Education; Family planning; Fertility; Foreign direct investment; Health; Health policy; HIV/AIDS; Human capital; Income; Instrumental variables; Investment; Labour productivity; Labour supply; Learning; Life expectancy; Morbidity; Mortality; Population growth; Population health; Retirement; Savings; School attendance; Social security; Value of life; Well-being

JEL Classification

I1

Population health and a high level of income go hand in hand. Higher incomes promote better health through improved nutrition, better access to safe water and sanitation, and increased ability to purchase more and better quality health care. There is also, however, an effect of health on income. This can work through several mechanisms (Bloom and Canning 2000). The first is the role of health in labour productivity. Healthy workers lose less time from work due to ill health and are more productive when working. The second is the effect of health on education. Childhood health can have a direct effect on cognitive development and the ability to learn. In addition, because adult mortality and morbidity (sickness) can lower the prospective returns to investments in schooling, improving adult health can raise the incentives to invest in education. The third is the effect of health on savings. A longer prospective lifespan can increase the incentive to save for retirement, generating higher levels of saving and wealth, and a healthy workforce can increase the incentives for business investment. We examine the evidence for these mechanisms and find that there are potentially large effects of health on economic outcomes at both individual and macroeconomic levels.

Improved population health has a large impact on population numbers and age structure, and we examine the economic implications of this induced demographic change. The global population explosion of the nineteenth and twentieth centuries was caused not by a rise in fertility but by a fall in mortality. Lower mortality and improved survival rates increased population numbers, but also led to significant increases in the number of young people since the largest improvements in mortality are initially in infant mortality rates. In the long run, reductions in infant mortality lead to a fall in desired fertility, creating a one-time baby-boom cohort. As this large cohort ages, the resultant changes in population age structure can have significant economic implications.

The issue of population health and economic outcomes is particularly acute in sub-Saharan Africa. This region has a high burden of tropical and other infectious disease, such as malaria, tuberculosis, and intestinal worms, and it also suffers from the HIV/AIDS pandemic. We examine the impact of this disease burden on the prospects for economic development in sub-Saharan Africa.

Although we focus on the economic implications of population health, there is clearly two-way causality as health is partly a consequence of income levels. Preston (1975) demonstrated a positive correlation between national income levels and life expectancy. One reason for this link is that higher income levels allow greater access to inputs that improve health, such as food, clean water and sanitation, education, and medical care. Fogel (2004) emphasizes the role of access to food while Deaton (2006) puts more weight on public health measures such as clean water and sanitation (see Cutler and Miller 2005). Cutler and McClellan (2001) examine the increasing contribution of medical care to health outcomes. Pritchett and Summers (1996) use the relationship between income levels and health to argue for an emphasis on economic growth in poor countries as a method of increasing population health. However, the findings of Easterly (1999) weaken this argument. Easterly finds that, although income levels and population health are closely related, the effect of changes in income on population health over reasonable time spans appears to be quite weak. By contrast, relatively inexpensive public health interventions and policies can have remarkable impacts on population health even in very poor countries. In practice, the major force behind health improvements has been improvements in health technologies and public health measures that prevent the spread of infectious disease, and not higher incomes (Cutler et al. 2006).

We examine the role of health as an instrument to generate economic wellbeing. However, any reasonable view of the contribution of health to human welfare would also include the direct welfare benefits of a long lifespan and good health. Estimates of the monetary value of life (as measured by the willingness to pay to avoid a small risk of death) are often very large (Viscusi and Aldy 2003). We can use these estimates of the value of life to compare the welfare improvements that have come about due to improvements in population health and the improvements due to economic growth and higher incomes. Such comparisons suggest that in many countries the value of health gains has been comparable to, or has even surpassed, the value of income gains (Nordhaus 2003; Becker et al. 2005).

Health as Human Capital

The idea of health as a form of human capital has a long history (for example, see Mushkin 1962). Grossman (1972) develops a model in which illness prevents work so that the cost of ill health is lost labour time. However, there may also be an effect of ill health on worker productivity in employment. A major difficulty in measuring the economic effect of health is the two-way causality between wealth and health (Smith 1999). Another difficulty is the lack of consensus on what is meant by health. Different studies use different health measures: self-assessments of health, biomarkers, medical records, limitations on physical functioning, and anthropometric measurements have all been used as health indicators. Each of these approaches may fail to provide a complete picture of an individual's health status, giving rise to a problem of measurement error. In addition, it is necessary to separate out the effect of investments in health from the effect of natural or genetic variation in health (Schultz 2005).

One solution to these problems in measuring the effect of health on worker productivity is to establish the causal paths in panel data through the use of timing of health shocks and income or wealth responses (for example, Adams et al. 2003). Case et al. (2005), controlling for parental influences and education, find that childhood health has a significant impact on adult health and earnings. Yet another approach to establishing causality is to use instrumental variables. For example, Schultz (2002) instruments adult height with childhood health and nutrition to argue that each centimeter gain in height due to improved inputs as a child in Ghana and Brazil leads to a wage increase of between 8 and 10% (Strauss and Thomas 1998, provide a survey of studies in this area).

Thomas and Frankenberg (2002) caution against drawing inferences from observational studies and instead advocate an experimental approach. A randomized experiment using iron supplementation to reduce iron deficiency anemia led to sizeable effects on worker productivity in Indonesia (Basta et al. 1979). Quasi-experiments can be used where it is possible to treat changes to health as if such changes were randomly generated. Bleakley (2003) considers the effects of the eradication of hookworm and malaria in the United States in the 1910s and 1920s. These diseases were pandemic in many counties of the American South prior to eradication. Bleakley, controlling for normal wage gains in areas that were not infected, shows that children not exposed to these diseases due to their eradication had improved incomes as adults relative to those born before eradication.

This body of research on health and human capital generally supports the idea that health affects worker productivity. However, it lacks a good appreciation of which types of health intervention are most important and what rate of return can be achieved by investing in health as a form of human capital. In many developing countries, relatively inexpensive activities designed to prevent the spread of infectious disease (for example, vaccination) can increase population health at low cost, suggesting that even modest income gains from health will generate very high rates of return. By comparison, treating chronic non-infectious disease in developed countries is often costly. There is evidence that susceptibility to chronic disease in later life is determined by health and nutrition as a fetus and in infancy (Barker 1992; Behrman and Rosenzweig 2004), suggesting that early health investments are crucial for adult productivity.

Health and Education

Education is widely agreed to affect economic outcomes, and health affects education through two mechanisms. The first is the effect of better child health on school attendance, cognitive ability, and learning. Bleakley (2003) finds that deworming of children in the American South had an effect on their educational achievements while in school. Miguel and Kremer (2004) find that deworming of children in Kenya increased school attendance.

The second mechanism is the effect of lower mortality and a longer prospective lifespan on increasing incentives to invest in human capital. This effect occurs for the individual for whom the benefits of education are now greater (Kalemli-Ozcan et al. 2000). In addition, lower infant mortality may encourage parents to invest more resources in fewer children, leading to low fertility but high levels of human capital investment in each child (Kalemli-Ozcan 2002). Evidence for this effect is limited, though Bils and Klenow (2000) do find an effect of life expectancy on investments in education at the national level.

Health and Saving

Poor health affects both the ability to save and the impetus to save. Sickness can have a large effect on out-of-pocket medical expenses, which can reduce current and accumulated household savings. This occurs in developed countries (Smith 1999) but is of particular concern in developing countries where families may be thrown into poverty if productive assets such as land or animals must be sold to pay for medical expenses.

Because poor health tends to be associated with a short lifespan, increasing population health and expected longevity will have an effect on the planning horizon and will influence life-cycle behaviour. With a fixed retirement age, a longer lifespan elicits greater savings for retirement. Blanchard (1985) considers the theoretical effect of a longer lifespan in a macroeconomic model. Hurd et al. (1998) find that increased expectation of longevity leads to greater wealth-holding at the household level in the United States. Bloom et al. (2003) find an effect of life expectancy on national savings, using cross-country data. Lee et al. (2000) argue that rising life expectancy can account for the boom in savings in Taiwan since the 1960s. But the effect of a longer lifespan need not be increased saving for retirement; people could instead choose to work longer. The behavioural response to longer lifespans depends on social security arrangements and retirement incentives (Bloom et al. 2007).

In a life-cycle model with a stable age structure and no population growth or economic growth, the dissaving of the old will exactly match the saving of the young at any level of life expectancy. This suggests that the aggregate effect of longer lifespans on savings is temporary and occurs when life expectancy rises. In the long run, the high savings rates of the working age population will be off set by the dissaving of a large cohort of elderly.

An effect on saving may lead to higher investment if capital markets are not perfectly open. In addition, a healthy population and workforce may increase productivity and encourage foreign direct investment (Alsan et al. 2006).

Health and Demography

Improvements in health and decreases in mortality rates can catalyse a transition from high to low rates of fertility and mortality – the 'demographic transition' (Lee 2003). Population growth is the difference between birth and death rates (ignoring migration) and the global population explosion in the twentieth century is attributable to improvements in health and falling death rates. In developing countries, health advances tend to lower infant and child mortality rates, leading initially to a surge in the number of children. Reduced infant mortality, increased numbers of surviving children, and rising wages for women can lower desired fertility (see Schultz 1997) leading to smaller cohorts of children in future generations. Better access to family planning can also help couples achieve match more closely their fertility desires and realizations. This process creates a 'baby boom' generation that is larger than both preceding and succeeding cohorts. Subsequent health improvements tend primarily to affect the elderly, reducing old-age mortality and lengthening the lifespan.

In many theoretical models a population explosion reduces income per capita by putting pressure on scarce resources and by diluting the capital–labour ratio. In these models population declines spur economic growth in per capita terms. For example, the very high death rates, and decline in population, due to the Black Death in fourteenth century Europe appear to have caused a shortage of labour, leading to a rise in wages and the breakdown of the feudal labour system (Herlihy 1997). However, in modern populations there appears to be little connection between overall population growth and economic growth; indeed the twentieth century saw both a population explosion and substantial rises in income levels.

Although it is difficult to find significant effects of overall population growth on economic growth, it is possible to consider the components of population growth separately. High birth and low death rates both generate population growth, but seem to have quite different effects on economic growth (Bloom and Freeman 1988; Kelley and Schmidt 1995). This may be because, while both forces increase population numbers, they affect the age structure quite differently. The effect of changing age structure due to a baby boom has large effects as the baby boomers enter the workforce and then as they eventually retire. While the baby boomers are of working age, economic growth may be spurred by a 'demographic dividend' if the baby boom generation can be productively employed. Bloom et al. (2004) find that the demographic dividend increases the potential labour supply but its effect on economic growth depends on the policy environment.

There is a worry that health improvements and population aging will lead to high dependency rates and a slowdown in economic growth. In addition to longer lifespans, however, we are seeing a compression of morbidity; the period of sickness towards the end of life is falling as a proportion of overall lifespan (Fries 1980, 2003). The idea that old-age dependency starts at 65 is essentially a result of social security retirement arrangements (Gruber and Wise 1998) and healthy aging means that physical dependency now often occurs at much later ages.

Health and Economic Growth

In growth models, population health is usually taken to be life expectancy, or some other mortality measure, as opposed to the morbidity measured used at the individual level. This disjunction can be bridged by assuming a one-to-one relationship between mortality and morbidity rates in a population; however it is not clear that such a relationship holds, making comparison of the macroeconomic relationship and microeconomic relationships difficult. In addition, calculating life expectancy requires age-specific mortality rates that are unavailable for many developing countries and published life-expectancy figures from the World Bank and United Nations are often constructed from quite incomplete raw data (Bos et al. 1992). There is a need to improve our measures of population health and to expand them to measures that correspond to morbidity and not just mortality.

The effect of health on individual productivity implies a relationship between population health and aggregate output. Shastry and Weil (2003) calibrate a production function model of aggregate output using microeconomic estimates of the return to health. They find that cross-country gaps in income levels can be explained in part by differential levels of physical capital, education, and health, with these three factors being roughly equal in terms of their contribution to differences in income levels. (A little over half of cross-country income gaps are explained by these factors; the remainder of the gap is ascribed to differences in total factor productivity.)

Another approach estimates the effect of population health on economic growth. Estimating the effect of the current level of population health on current income levels is subject to the problem of reverse causality; income also affects health. One way around this problem is to look at the effect of population health on subsequent economic growth, arguing that the timing can determine the direction of causality. This requires the absence of reverse causality through an expectation effect (so that current health is not caused by expected future economic growth).

Growth regressions show that the initial levels of population health are a significant predictor of future economic growth (Bloom et al. 2004, provide a survey of this literature). Sala-i-Martin et al. (2004) find that the predictive power of health (as measured by life expectancy and malaria prevalence) is robust to the specification of the growth regression. Bhargava et al. (2001) argue that the effect of health on economic growth is larger in developing countries than in developed countries.

While population health measures are highly predictive of future economic growth, there is a debate about how to interpret the link. The health effect could be interpreted as the macroeconomic counterpart of the worker productivity effect found in individuals. However, Acemoglu et al. (2003) argue that health differences are not large enough to account for much of the cross-country difference in incomes, and that the variations in political, economic and social institutions are more central factors. They argue that health does not have a direct effect on growth, but serves in growth regressions as a proxy for the pattern of European settlement, which was more successful in countries with a low burden of infectious disease.

Even if a causal interpretation of the effect of health on individual productivity and economic growth is accepted, the argument for using health as an input depends on there being low-cost health interventions that can increase population health without first having a high income level. There are, however, a large number of such interventions that can be implanted (Commission on Macroeconomics and Health 2001).

Tropical Disease and HIV/AIDS

Sub-Saharan Africa suffers from poor health due to the widespread presence of tropical disease. Malaria and tuberculosis cause high illness and death rates, while parasitic diseases such as schistosomiasis and intestinal worms can cause anemia and reduced energy levels and productivity. In addition to these tropical diseases, the high prevalence of HIV/AIDS is causing life expectancy to decline dramatically in many countries in the region. Poor health status is one cause of sub-Saharan Africa's economic stagnation (Bloom and Sachs 1998). Malaria appears to have an effect on economic growth over and above that created through higher mortality, suggesting that its effects on productivity with a given mortality burden are greater than other diseases (Gallup and Sachs 2001).

Although HIV/AIDS has increased mortality rates dramatically, its impact on income per capita is unclear. HIV/AIDS is associated with high mortality but the period of sickness before death is relatively short. This mutes the worker productivity effects of the disease. Bloom and Mahal (1997) find that HIV/AIDS does not seem to lower the growth rate of income per capita; lower output is matched by lower population numbers due to high death rates. Young (2005) goes further and argues that AIDS mortality reduces fertility significantly, and that this will lower population pressure and increase the income per capita of the survivors of the pandemic in South Africa.

Many authors, however, argue that AIDS mortality has significant indirect effects that will reduce economic growth in the long term. Deaths from HIV/AIDS are concentrated among young adult men and women, leading to a higher dependency ratio. Bell et al. (2004) argue that the creation of a generation of AIDS orphans may lead to lack of care and education for children and to low productivity in the future. This effect may be compounded by fatalism induced by high AIDS mortality and shortened expected lifespan, which reduce the return to education. The high level of stigma associated with HIV/AIDS can reduce trust in the community, while high mortality and the strains imposed by extreme ill health before death can weaken families, community groups, firms, and government agencies, with long-term consequences for social capital (Haacker 2004).

It is important to remember that income per capita is not a complete measure of welfare. Resources devoted to preventing and treating HIV/AIDS are part of measured income but reduce consumption of other goods, reducing welfare even as measured GDP per capita may remain steady (Canning 2006). A more comprehensive welfare measure that included the welfare gain derived from a long lifespan, as well as annual income, would show a large welfare reduction due to HIV/AIDS (Crafts and Haacker 2004). The main welfare effect of HIV/AIDS is the sickness and death of its victims and the impact of these on the victims' families; the effect on the average income level of the survivors is decidedly secondary.

See Also

▶ Child Health and Mortality
▶ Demographic Transition
▶ Education in Developing Countries
▶ Fertility in Developing Countries
▶ Health Outcomes (Economic Determinants)
▶ Human Capital, Fertility and Growth
▶ Malthusian Economy
▶ Population Ageing
▶ Retirement
▶ Returns to Schooling
▶ Value of Life

Bibliography

Acemoglu, D., S. Johnson, and J. Robinson. 2003. Disease and development in historical perspective. *Journal of the European Economic Association, Papers and Proceedings* 1: 397–405.

Adams, P., M.D. Hurd, D.L. McFadden, A. Merrill, and T. Ribeiro. 2003. Healthy, wealthy, and wise? Tests for direct causal paths between health and socioeconomic status. *Journal of Econometrics* 112: 3–56.

Alsan, M., D.E. Bloom, and D. Canning. 2006. The effect of population health on foreign direct investment inflows to low- and middle-income countries. *World Development* 34: 613–630.

Barker, D.J.P. 1992. *The fetal and infant origins of adult disease*. London: BMJ Books.

Basta, S., K. Soekirman, and N. Scrimshaw. 1979. Iron deficiency anemia and productivity of adult males in Indonesia. *American Journal of Clinical Nutrition* 32: 916–925.

Becker, G.S., T.J. Philipson, and R.R. Soares. 2005. The quantity of life and the evolution of world inequality. *American Economic Review* 95: 277–291.

Behrman, J.R., and M.R. Rosenzweig. 2004. The returns to birthweight. *Review of Economics and Statistics* 86: 586–601.

Bell, C., S. Devarajan, and H. Gersbach. 2004. Thinking about the long-run economic costs of AIDS. In *The Macroeconomics of HIV/AIDS*, ed. M. Haacker. Washington, DC: International Monetary Fund.

Bhargava, A., D. Jamison, L. Lau, and C. Murray. 2001. Modeling the effects of health on economic growth. *Journal of Health Economics* 20: 423–440.

Bils, M., and P.J. Klenow. 2000. Does schooling cause growth? *American Economic Review* 90: 1160–1183.

Blanchard, O.J. 1985. Debt, deficits, and finite horizons. *Journal of Political Economy* 93: 223–247.

Bleakley, H. 2003. Disease and development: Evidence from the American south. *Journal of the European Economic Association* 1: 376–386.

Bloom, D.E., and D. Canning. 2000. The health and wealth of nations. *Science* 287: 1207–1208.

Bloom, D.E., and R.B. Freeman. 1988. Economic development and the timing and components of population growth. *Journal of Policy Modeling* 10(1): 57–82.

Bloom, D.E., and A.S. Mahal. 1997. Does the AIDS epidemic threaten economic growth? *Journal of Econometrics* 77: 105–124.

Bloom, D.E., and J. Sachs. 1998. Geography, demography, and economic growth in Africa. *Brookings Papers on Economic Activity* 1998(2): 207–273.

Bloom, D.E., D. Canning, and B. Graham. 2003. Longevity and life-cycle savings. *Scandinavian Journal of Economics* 105: 319–338.

Bloom, D.E., D. Canning, and J. Sevilla. 2004. The effect of health on economic growth: A production function approach. *World Development* 32: 1–13.

Bloom, D.E., D. Canning, R. Mansfield, and M. Moore. 2007. Demographic change, social security systems, and savings. *Journal of Monetary Economics* 54: 92–114.

Bos, E., M.T. Vu, P.W. Stephens, and W. Patience. 1992. *Policy research working paper series 851. Sources of world bank estimates of current mortality rates*. Washington, DC: World Bank.

Canning, D. 2006. The economics of HIV/AIDS in Low-income countries: The case for prevention. *Journal of Economic Perspectives* 20: 121–142.

Case, A., A. Fertig, and C. Paxson. 2005. The lasting impact of childhood health and circumstance. *Journal of Health Economics* 24: 365–389.

Commission on Macroeconomics and Health. 2001. *Macroeconomics and health: Investing in health for economic development*. Geneva: World Health Organization.

Crafts, N., and M. Haacker. 2004. Welfare implications of HIV/AIDS. In *The Macroeconomics of HIV/AIDS*, ed. M. Haacker. Washington, DC: International Monetary Fund.

Cutler, D.M., and M. McClellan. 2001. Productivity change in health care. *American Economic Review* 91: 281–286.

Cutler, D.M., and G. Miller. 2005. The role of public health improvements in health advances: The twentieth-century United States. *Demography* 42: 1–22.

Cutler, D.M., A.S. Deaton, and A. Lleras-Muney. 2006. The determinants of mortality. *Journal of Economic Perspectives* 20(3): 71–96.

Deaton, A. 2006. The great escape: A review essay on Fogel's the escape from hunger and premature death, 1700–2100. *Journal of Economic Literature* 44: 106–114.

Easterly, W. 1999. Life during growth. *Journal of Economic Growth* 4: 239–276.

Fogel, R.W. 2004. *The escape from hunger and premature death, 1700–2100: Europe, America, and the third world*. Cambridge, MA: Cambridge University Press.

Fries, J.F. 1980. Aging, natural death, and the compression of morbidity. *New England Journal of Medicine* 303: 130–135.

Fries, J.F. 2003. Measuring and monitoring success in compressing morbidity. *Annals of Internal Medicine* 139: 455–459.

Gallup, J.L., and J.D. Sachs. 2001. The economic burden of malaria. *American Journal of Tropical Medicine and Hygiene* 64(1, 2 Supplement): 85–96.

Grossman, M. 1972. On the concept of health capital and the demand for health. *Journal of Political Economy* 80: 223–255.

Gruber, J., and D. Wise. 1998. Social security and retirement: An international comparison. *American Economic Review* 88(2): 158–163.

Haacker, M. 2004. HIV/AIDS: The impact on the social fabric and the economy. In *The macroeconomics of HIV/AIDS*, ed. M. Haacker. Washington, DC: International Monetary Fund.

Herlihy, D. 1997. *The Black Death and the transformation of the west*. Cambridge, MA: Harvard University Press.

Hurd, M., D. McFadden, and L. Gan. 1998. Subjective survival curves and life-cycle behavior. In *Inquiries in the economics of aging*, ed. D. Wise. Chicago: University of Chicago Press.

Kalemli-Ozcan, S. 2002. Does mortality decline promote economic growth? *Journal of Economic Growth* 7: 411–439.

Kalemli-Ozcan, S., H.E. Ryder, and D.N. Weil. 2000. Mortality decline, human capital investment, and economic growth. *Journal of Development Economics* 62: 1–23.

Kelley, A.C., and R.M. Schmidt. 1995. Aggregate population and economic growth correlations: The role of the components of demographic change. *Demography* 32: 543–555.

Lee, R. 2003. The demographic transition: Three centuries of fundamental change. *Journal of Economic Perspectives* 17(4): 167–190.

Lee, R., A. Mason, and T. Miller. 2000. Life cycle saving and the demographic transition: The case of Taiwan. *Population and Development Review* 26(Supplement): 194–219.

Miguel, E., and M. Kremer. 2004. Worms: Identifying impacts on education and health in the presence of treatment externalities. *Econometrica* 72: 159–217.

Mushkin, S.J. 1962. Health as an investment. *Journal of Political Economy* 70(5): 129–157.

Nordhaus, W. 2003. The health of nations: The contribution of improved health to living standards. In *Measuring the gains from medical research: An economic approach*, ed. K.H. Murphy and R.H. Topel. Chicago: University of Chicago Press.

Preston, S. 1975. The changing relation between mortality and level of economic development. *Population Studies* 29: 231–248.

Pritchett, L., and L. Summers. 1996. Wealthier is healthier. *Journal of Human Resources* 31: 841–868.

Sala-i-Martin, X., G. Doppelhofer, and R.I. Miller. 2004. Determinants of long-term growth: A Bayesian averaging of classical estimates (BACE) approach. *American Economic Review* 94: 813–835.

Schultz, T.P. 1997. The demand for children in low income countries. In *Handbook of population and family economics*, vol. 1A, ed. M.R. Rosenzweig and O. Stark. Amsterdam: North-Holland.

Schultz, T.P. 2002. Wage gains associated with height as a form of human capital. *American Economic Review: Papers and Proceedings* 92: 349–353.

Schultz, T.P. 2005. Productive benefits of health: Evidence from low income countries. In *Health and economic growth: Findings and policy implications*, ed. G. Lopez-Casasnovas, B. Riveras, and L. Currais. Cambridge, MA: MIT Press.

Shastry, G.K., and D.N. Weil. 2003. How much of cross-country income variation is explained by health? *Journal of the European Economic Association* 1: 387–396.

Smith, J.P. 1999. Healthy bodies and thick wallets: The dual relation between health and economic status. *Journal of Economic Perspectives* 13(2): 145–166.

Strauss, J., and D. Thomas. 1998. Health, nutrition and economic development. *Journal of Economic Literature* 36: 766–817.

Thomas, D., and E. Frankenberg. 2002. Health, nutrition and prosperity: A microeconomic perspective. *Bulletin of the World Health Organization* 80(2): 106–113.

Viscusi, W.K., and J.E. Aldy. 2003. The value of a statistical life: A critical review of market estimates from around the world. *Journal of Risk and Uncertainty* 27: 5–76.

Young, A. 2005. The gift of the dying: The tragedy of AIDS and the welfare of future African generations. *Quarterly Journal of Economics* 120: 243–266.

Portfolio Analysis

Nils H. Hakansson

Many observers trace the beginnings of modern financial investment theory to the pioneering article of Markowitz (1952), published only a third of a century ago. This is not surprising in view of the dominant position that the mean-variance approach to portfolio choice analysed by Markowitz has attained in the last two decades, particularly in empirical studies. Financial investment theory under uncertainty goes well beyond this particular model, however, and somewhat further back in time as well. This entry will first examine the pure portfolio model, both the single-period and the intertemporal varieties. It will then turn to consumptioninvestment formulations.

Pure Portfolio Analysis

Single-Period Models

Even though the mean-variance model 'dominates' single-period analysis, it will be expedient to begin with the approach which is a direct application of the theory of rational choice, also known as expected utility portfolio models.

The Expected Utility Approach

The investor, starting the period with initial capital $w_0 > 0$, is assumed to have preferences that are rational (in the von Neumann and Morgenstern (1944) sense) with respect to end-of-period distributions of wealth and therefore representable by a utility function, u, defined on end-ofperiod wealth w. Thus, the investor's problem is to maximize $E[u(w)]$, where E denotes the expectation operator. Letting r_i denote the (generally random) return per unit of investment in opportunity i and z_i the amount (to be) invested in opportunity (asset, security) i, $i = 1, \ldots, m$, we obtain

$$w = \sum_i z_i(1 + r_i), \qquad \sum_i z_i = w_0,$$

where the second expression is the budget constraint. Solving the second expression for z_1 and inserting the result in the first equality, the investor's problem becomes

$$\text{P1}: \max_{z_2, \ldots, z_m} E\left\{ u\left[\sum_{i=2}^{m} (r_i - r_1) z_i + w_0(1 + r_1) \right] \right\} \tag{1}$$

subject to

$$\text{miscellaneous constraints.} \tag{2}$$

At this point, several remarks are in order. First, in our expression for w we have implicitly assumed a perfect market, that is an absence of transaction costs and taxes, perfect divisibility, a competitive securities market, constant returns to scale, and that the investor has full use of the proceeds from short sales (negative holdings). These assumptions are standard and will be maintained throughout. Second, when some security is risk-free over the holding period, it carries the subscript $i = 1$ above; in this case, the first $m - 1$ terms in (1) represents the excess earned (over and above what an entirely risk-free portfolio would have provided) on the risky holdings (this excess may of course be negative). Third, it is usually assumed (quite innocuously from an empirical viewpoint) that the investor prefers more to less and is averse to risk, that is that

$$u' > 0, u'' < 0. \tag{3}$$

Finally, the constraints (2) usually represent institutional and/or self-imposed barriers on borrowing (e.g. margin requirements), on short positions, and on solvency (such as $\Pr\{w > 0\} = 1$).

The solution to P1 is usually denoted $z^*(w_0) = z_2^*(w_0), \ldots, Z_m^*(w_0)$. It exists under various innocent conditions: one set imposes bounded returns on the available securities, 'no-easymoney', and a solvency constraint. The no-arbitrage or no-easy-money condition precludes both a payoff $w \geq 0$, where $\Pr\{w > 0\} > 0$, from a nonpositive net investment, as well as a payoff $w = 0$ from a negative net investment. Given existence, the second part of (3) (strict concavity of u) implies that the optimal payoff distribution w^* (though not necessarily the optimal portfolio z^*) will be unique.

Define $a(w)$ (the absolute risk aversion function) and $r(w)$ (the relative risk aversion function) by

$$a(w) \equiv -u''(w)/u'(w), \qquad r(w) \equiv wa(w).$$

Arrow (1965) demonstrated that if $E[r_2] > r$,

$$a'(w) \gtreqless 0 \Rightarrow \frac{dz_2^*}{dw_0} \lesseqgtr 0$$

when there are only two assets available, one risky and one risk-free. While the result does not extend in general to the case of many risky assets (Cass and Stiglitz 1972), the empirical observation that a given portfolio of risky assets is overwhelmingly treated as a normal (as opposed to inferior) good lends strong support to the notion that the preferences of the great majority of investors have the property

$$a'(w) < 0 \tag{4}$$

in addition to those given in (3). Beyond this, however, we have little to say about investors' preference functions with respect to wealth.

Since properties (3) and (4) leave much room for individuality, there is rather little one can say in general about the solution to P1 – except that the optimal portfolio will be well diversified. This observation was probably first made in a scholarly context by Bernoulli (1738) in his advocacy of the logarithmic measure of welfare.

There are, however, two cases of special interest. One is the case in which the optimal investment policy is *proportional* to initial capital. This occurs if and only if utility is a member of the family of power functions (the isoelastic family), that is

$$u(w) = \begin{cases} -w^\gamma, & \gamma < 0 \\ \ln w, & (\gamma = 0) \\ w^\gamma, & 0 < \gamma < 1 \end{cases}, \tag{5}$$

which in turn implies, and is implied by, constant relative risk aversion. [For the family above, $r(w) = 1 - \gamma$.] The optimal policy is now of the form

$$z_i^*(w_0) = x_{iy}^* w_0, \qquad \text{all } i, \gamma, \tag{6}$$

where the x_{iy}^* are constants corresponding to the proportions to be invested in the various assets.

A second special case is that of *linear* optimal investment policies (of which (6) is obviously a special case). This occurs, *assuming a risk-free asset or portfolio is available*, if and only if preferences exhibit linear risk tolerance $[a(w)^{-1}$ is linear] or, equivalently, hyperbolic absolute risk aversion, that is

$$u(w) = \begin{cases} \gamma^{-1}(w+\phi)^\gamma & \gamma < 1 & a' < 0 & \text{(7a)} \\ -(\phi - w)^\gamma & \gamma > 0, \ \phi \text{large} & a' > 0 & \text{(7b)} \\ -\exp\{\phi w\} & \phi < 0 & a = 0. & \text{(7c)} \end{cases}$$

The optimal policies are given, in the three cases, by

$$z_i^*(w_0) = \begin{cases} x_{iy}^*\left(w_0 + \dfrac{\phi}{1+r_1}\right) & & \text{(8a)} \\ x_{iy}^*\left(\dfrac{\phi}{1+r_1} - w_0\right) & i = 2, \ldots, m & \text{(8b)} \\ \text{a constant } (\phi) & & \text{(8c)} \end{cases}$$

and are said to exhibit the *separation* property. This name derives from the fact that the *mix* of risky assets (the ratio of $Z_i^*(w_0)/z_j^*(w_0)$ any i, j $\geq$ 2) is independent of initial wealth w_0 (it is also independent of the preference parameter ϕ). *In the absence of a risk-free asset or portfolio*, separation obtains only for quadratic utility, or when $\gamma = 2$ in (7b). Thus we have the remarkable observation that, for arbitrary return distributions, two individuals of differing initial wealth levels would be willing to delegate the choice of risky asset proportions to the same mutual fund only if they share probability beliefs *and either* a risk-free portfolio is available and both individuals' preference functions belong to either (7a) or (7b) with a common γ or to (7c), *or* both investors have quadratic utility. When it comes to risky investments, individuality runs strong indeed!

Separation based on return distributions rather than preferences can also occur but only under highly restrictive assumptions (Ross 1978). The most noteworthy case is when returns are normally distributed, which is discussed in the next section.

The Mean-Variance Approach

The essence of the mean variance model is that more expected return is preferred to less and that less variance of return is preferred to more, *ceteris paribus*. In addition, it is usually assumed that indifference curves in standard deviation-mean space are convex. Since the return r on a portfolio is $w/w_0 - 1$, we obtain, defining x_i as the fraction of w_0 invested in opportunity i or $x_{\mathrm{i}} \equiv z_{\mathrm{i}}/w_0$ and using (1),

$$r(x) = \sum_{i=2}^{m} (r_i - r_1)x_i + 1 + r_1.$$

More formally, the mean-variance approach can thus be viewed as postulating a preference function $f(E[r], V[r])$, where $V[r]$ is the variance of r, such that

$$\frac{\partial f}{\partial E} > 0, \qquad \frac{\partial f}{\partial V} < 0, \qquad \left.\frac{\mathrm{d}^2 E}{\mathrm{d}\left(\sqrt{V}\right)^2}\right|_{f=f_0} > 0. \quad (9)$$

The first two properties of (7) provide the basis for the central notion of mean-variance dominance: return distribution r_i is said to MV-dominate distribution r_k if and only if

$$E_i \geq E_k, \qquad V_i \leq V_k$$

and at least one inequality is strict. Given the set of feasible portfolios, dominated portfolios are referred to as *inefficient* and nondominated portfolios as *efficient*. The first two properties of (7) thus generate a partial ordering of payoff distributions in a manner similar to that of the various stochastic dominance criteria.

In the absence of a risk-free asset or portfolio, $E[r]$ is (except in pathological cases) a strictly concave function of $\sigma[r](= \sqrt{V[r]})$ for the set of efficient portfolios. In the presence of a risk-free asset, the expected return of any efficient portfolios p is given by the linear equation

$$E\left[r_p\right] = r_1 + \frac{E[A] - r_1}{\sigma[A]}\sigma\left[r_p\right],$$

where A is the one portfolio composed solely of risky assets that is efficient. In other words, all efficient portfolios are combinations of the risk-free asset and portfolio A, that is the separation property holds.

As noted, Markowitz is viewed as the originator of mean-variance portfolio theory, although Tobin (1958) also made important early contributions. However, the mean variance approach itself has three other independent and rather interesting origins. Marschak (1951), using a Taylor series expansion as an approximation to the expected utility of return, obtained, on the basis of the first three terms, the expression

$$E[r] - b(E[r])^2 - bV[r], \qquad b > 0,$$

which is an eligible form of the mean-variance function $f(E, V)$. Roy (1952) argued for maximizing the probability of exceeding some disaster level d, or the criterion

$$\max \Pr\{r > d\}.$$

Applying Chebychev's inequality, he obtained the operational expression

$$\max_x \frac{E[r(x)] - d}{\sigma[r(x)]}.$$

which clearly captures the essence of the mean variance framework. Finally, Freund (1956), assuming negative exponential utility [see 7(c)] and normally distributed returns, obtained

$$E[u(w)] = -\exp\left\{k\left(E[w] + \frac{k}{2}V[w]\right)\right\}, k < 0,$$

where, upon optimization, each permissible value of k implies a mean-variance efficient solution.

The mean-variance model is consistent with the expected utility criterion in two principal cases. First, under arbitrary return distributions, utility must be quadratic [$u(w) = w - bw^2, b > 0$], which unfortunately implies $u' \leqslant 0$ for $w \geqslant b/2$ and that risky assets are inferior goods (see 8b). Second, when returns are normally distributed, consistency occurs for that subset of preferences for which the expected utility integral exists (a necessary condition for this is that $u(w)$ is defined on the whole real line – this excludes the family (7a), for example).

Although normally distributed returns are a poor approximation of actual returns in a world of limited liability, and quadratic utility leaves much to be desired, the mean-variance model is by far the most widely used. This appears to be attributable to three principal properties. First, MV-efficient portfolios are (like the portfolios of risk averse expected utility maximizers) well diversified. Second, the MV-model makes more modest input demands and is computationally much simpler than the (non-quadratic) expected utility models. (What business person would appreciate the advice that (s)he maximize expected utility?) Finally, the normality assumption appears to provide a reasonable approximation of the returns for well diversified portfolios in many cases, and the quadratic function, over a limited range, is often a satisfactory approximation to an arbitrary utility function.

Multi-Period Models

This section addresses the type of models in which a large number of sequential portfolio choices is of the essence. We shall therefore employ the subscript t to denote period t; w_t represents wealth at the end of period t. The returns r_{it} will be assumed to be independent with respect to t (but not i).

The Long-Run Growth Model

Let $R_t(xt) \equiv 1 + r_t(x_t)$; R_t is now called the wealth relative for period t. Thus, under full reinvestment of the previous period's payoffs.

$$\begin{aligned} w_t &= w_0 R_1(x_1)\ldots R_t(x_t) \\ &= w_0 \exp\left\{\sum_{n=1}^{t} \ln R_n(x_n)\right\}, \end{aligned}$$

where we assume that $R_t(x_t) \geq 0$, all t. Letting

$$G_t(\langle w_t \rangle) \equiv \sum_{n=1}^{t} \ln R_n(x_n)/t \qquad (10)$$

and observing that the variates $1\text{n}\ R_1$, $1\text{n}\ R_2$, ... (under mild restrictions) obey the law of large numbers, we obtain

$$w_t \to \begin{cases} 0 & \text{if } E[G_t] \leqslant \delta < 0 \\ \infty & \text{if } E[G_t] \geqslant \delta > 0 \end{cases} \quad t \geqslant T, T \text{ large.} \qquad (11)$$

Thus, it is the expectations of the logs of the wealth relatives which are the principal determinants of what happens to your capital over the long haul.

In view of (9), it is natural to think of maximizing the expectation of G since this almost surely leads to more capital in the long run than any other (significantly different) strategy. To do this, it is necessary and sufficient to

$$\max_{x_t} E[\ln R_t(x_t)], \quad \text{each } t, \qquad (12)$$

that is to solve (10) one period at a time. Note that (10) is equivalent to maximizing the geometric mean of Rt in each period. This model appears to have been independently discovered by Williams (1936), Kelly (1956), Latané (1959), and Breiman (1960).

The long-run growth model has several noteworthy properties. First, the decision rule (10) implies, and is implied by, logarithmic utility of wealth in each period. Thus, it is inconsistent with all (significantly) different preferences (including the mean-variance model). In other words, almost surely having more capital does not imply higher expected utility (or conversely). Various writers have on occasion been confused on this point.

Second, the ‘growth-optimal’ investment policy is not only proportional to initial wealth but (10) implies that it is *myopic*, that is independent of the return distributions beyond the current period (this is true even under returns that are weakly dependent over time). Finally, with relative risk aversion equal to 1, the model tells us that to do well in the long run in terms of capital accumulation, one must be averse to risk; furthermore, both greater and smaller risk aversion almost surely leave one with less capital than logarithmic risk aversion.

Terminal Utility Models

Now consider the case in which the investor’s preferences for wealth w at some (distant) terminal point in time h are represented by the utility function $U_h(w_h)$. Letting w_n be the investor’s wealth with n periods to go, we obtain, under full reinvestment of each period’s proceeds,

$$w_{n-1}(z_n) = \sum_{i=2}^{m} (r_{in} - r_{1n})z_{in} + w_n(1 + r_{1n}),\quad n = 1, 2, \ldots$$

where, for convenience, we set $h = 0$. Defining $U_n(w_n)$ as the maximum expected utility obtainable with w_n, we obtain the recursive equation

$$U_n(w_n) \equiv \max_{z_n} E\{U_{n-1}[w_{n-1}(z_n)]\},\quad n = 1, 2, \ldots \tag{13}$$

Consequently, $U_n(w_n)$ is the derived or induced utility of wealth with n periods to go.

The conditions for the existence of a solution to system (11) are the same as for the singleperiod model; when U_0 has properties (3), so do the induced functions $U_1, \ldots, U_n$. In general, $U_n(w_n)$ depends on all of the inputs: U_0, the joint distribution functions $F_1(r_1), \ldots, F_n(r_n)$, and the interest rates $r_{11}, \ldots, r_{1n}$. There are, however, two special cases. First, when $U_0(w_0)$ belongs to class (5), U_n becomes a positive linear transformation of U_0 so that in effect

$$U_n(w) = U_0(w),\quad n = 1, 2, \ldots$$

Consequently, the optimal investment policy $z_n^*(w_n)$ depends in this case only on the current periods inputs, $F_n(r_n)$ and r_{1n}, and is thus *myopic*. This was first shown by Mossin (1968).

The second special case occurs when interest rates follow a deterministic process. Then, when U_0 belongs to class (7a) with $\phi \leq 0$. U_n depends only on U_0 and $r_{11}, \ldots, r_{1n}$, which is called *partial myopia.* U_n and z_n^* are now given by

$$\begin{aligned} U_n(w_n) &= \gamma^{-1}(w_n + A_n)^\gamma, && \gamma < 1 \\ z_{in}^*(w_n) &= x_{in\gamma}^*(w_n + A_n), && i = 2, \ldots, m, \end{aligned}$$

where $A_n = \phi[(1 + r_{11})\ldots(1 + r_{1n})]^{-1}$. In the other cases of family (7), partial myopia occurs locally, that is for w_n greater than or equal to a (positive) lower bound.

The most interesting aspect of the terminal utility model, however, is a strong set of convergence results (see e.g. Hakansson 1974). Under very general conditions, we obtain from (11) that U_n converges to a member of the isoelastic family (5), that is

$$U_n(w_n) \to \frac{1}{\gamma} w_n^\gamma,\quad \text{some } \gamma < 1.$$

In addition,

$$z_n^*(w_n) \to x_{n\gamma}^* w_n.$$

Thus, we have the remarkable result that reinvesting individuals with distant horizons should follow an isoelastic investment policy independently of their terminal preferences as long as their horizon remains distant.

The Continuous-Time Model

Since transaction costs are zero under the perfect market assumption, it is natural to consider shorter and shorter periods between reinvestment decisions. In the limit, reinvestment takes place continuously. Assuming that the returns on risky assets can be described by diffusion processes, we obtain that optimal portfolios are mean-variance efficient in that the instantaneous variance is minimized for a given instantaneous expected return. The intuitive reason for this is that as the trading interval is shortened, the first two moments of the change in a security's price become more and more dominant (see Samuelson 1970). The optimal portfolios also exhibit the separation property – as if returns over very short periods were normally distributed. Over any fixed interval, however, payoff distributions are, due to the compounding effect, usually lognormal.

Consumption-Investment Analysis

In consumption-investment models, investment is merely a means to an end – future consumption and bequests. Thus, preferences are defined on consumption and bequest programmes, c_1, c_2, $\ldots$, c_n, b_n, where c_t is the *level* of consumption in period t and b_n the bequest at the end of the last period, assuming death occurs in period n. The utility of wealth is therefore not a primitive but must be induced or derived. Preferences may of course be conditional on n and depend on the environment s, in which case they may be written

$$U_{ns}(c_1 \ldots, c_n, b_n), \tag{14}$$

where it is usually assumed that the functions U_{ns} reflect a preference for more to less and are strictly concave. Commonly studied forms of (12) are those in which (12) is additive or multiplicative. When additive and state-independent, (12) may be written

$$u_1(c_1) + u_2(c_2) + \ldots + u_n(c_n) + g_n(b_n).$$

Wealth is now governed by the difference equation

$$\begin{aligned} w_{t+1} = {} & \sum_{i=2}^{m} (r_{its} - r_{1ts}) z_{it} \\ & + (w_t - c_t)(1 + r_{1ts}) + y_{ts}, \end{aligned}$$

where y_{ts} is employment income.

The simplest consumption–investment model is based on just two periods and can profitably be used to study such questions as, 'How does the investor respond to increasing investment risk?' Answer: any which way – see e.g. Rothschild and Stiglitz (1971). In multiperiod formulations, additional issues that must be addressed are the probabilistic nature of the investor's lifespan and the stochastic process obeyed by returns. Dynamic programming formulations of this problem become rather lengthy (see e.g. Hakansson 1970, 1971). The most general models posit a state-contingent opportunity set where the states obey a Markov process.

When preferences are either additive or multiplicative and $u_t(c_t)$ belongs to the family (7a) with $\phi \leq 0$, the separation property is preserved; when $\phi < 0$, $-\phi$ assumes the role of subsistence level. For the family (7b) and (7c), on the other hand, the non-negativity constraint on consumption is generally binding and poses insurmountable problems for the mean-variance model. However, as in the pure reinvestment model, mean-variance efficiency is restored by moving to a continuous-time formulation (Merton 1971).

Bibliography

Arrow, K. 1965. *Aspects of the theory of risk-bearing*. Yrjö Jahnsson Säätiö: Helsinki.

Bernoulli, D. 1738. Exposition of a new theory on the measurement of risk. Trans. by L. Sommer. *Econometrica* 22, January 1954, 23–36.

Breiman, L. 1960. Investment policies for expanding business optimal in a long-run sense. *Naval Logistics Quarterly.*

Cass, D., and J. Stiglitz. 1972. Risk aversion and wealth effects on portfolios with many assets. *Review of Economic Studies* 39(3): 331–354.

Freund, R. 1956. The introduction of risk into a programming model. *Econometrica* 24: 253–263.

Hakansson, N. 1970. Optimal investment and consumption strategies under risk for a class of utility functions. *Econometrica* 38(5): 587–607.

Hakansson, N. 1971. Optimal entrepreneurial decision in a completely stochastic environment. *Management Science: Theory* 17(7): 427–449.

Hakansson, N. 1974. Convergence to isoelastic utility and policy in multiperiod portfolio choice. *Journal of Financial Economics* 1(3): 20–24.

Kelly Jr., J.L. 1956. A new interpretation of information rate. *Bell System Technical Journal* 35: 917–925.
Latané, H. 1959. Criteria for choice among risky ventures. *Journal of Political Economy* 67: 144–155.
Markowitz, H. 1952. Portfolio selection. *Journal of Finance* 7(1): 77–91.
Marschak, J. 1951. Why 'should' statisticians and businessmen maximize 'moral expectation'? In *Proceedings of the second Berkeley symposium on mathematical statistics and probability*, ed. J. Neyman. Berkeley: University of California Press.
Merton, R. 1971. Optimum consumption and portfolio rules in a continuous time model. *Journal of Economic Theory* 3(4): 373–413.
Mossin, J. 1968. Optimal multiperiod portfolio policies. *Journal of Business* 41(April): 215–229.
Ross, S. 1978. Mutual fund separation in financial theory: The separating distributions. *Journal of Economic Theory* 17(2): 254–286.
Rothschild, M., and J. Stiglitz. 1971. Increasing risk II: Its economic consequences. *Journal of Economic Theory* 3(1): 66–84.
Roy, A. 1952. Safety first and the holding of assets. *Econometrica* 20: 431–449.
Samuelson, P.A. 1970. The fundamental approximation theorem of portfolio analysis in terms of means, variances, and higher moments. *Review of Economic Studies* 37(4): 537–542.
Tobin, J. 1958. Liquidity preference as behavior towards risk. *Review of Economic Studies* 25: 65–86.
von Neumann, J., and O. Morgenstern. 1944. *Theory of games and economic behavior*. Princeton: Princeton University Press.
Williams, J. 1936. Speculation and the carryover. *Quarterly Journal of Economics* 50: 436–455.

Portugal, Economics in

António Almodovar and José Luís Cardoso

Abstract

The development of economics in Portugal has been marked by intellectual curiosity coupled with pragmatism. Both characteristics are explained by the long-standing feeling that, although Portugal was lagging in terms of social and economic development, the situation could be overcome by means of an appropriate economic policy. This feeling motivated a continuing effort to find answers to economic and financial problems by careful analysis of other countries' experiences – both the principles discussed by economists and the policies eventually implemented by governments. Portuguese experience thus well illustrates the international diffusion of the ideas associated with different schools of economic thought.

Keywords

Austrian School; Corporatism; Free trade; Harrod–Domar growth model; Keynes, J. M.; Lausanne School; List, F.; Manchester School; Marginal revolution; Mercantilism; Neoclassical synthesis; Ophelimity; Physiocracy; Portugal, economics in; Protection; Quesnay, F.; Say, J.-B.; Smith, A.; Utility

JEL Classifications

B2

Mercantilism

The first interesting examples of a concern to establish first principles to explain economic reality emerged in Portugal in the 16th century. Extending the spirit of the Discoveries by the early Portuguese explorers to the scantily studied areas of economic knowledge, Portuguese authors showed a certain pioneering spirit. In contrast with former prejudices about the harmfulness of trade, commerce began to be considered as the principal cause of the wealth: commerce dynamically connected the different sectors of economic activity and brought individuals and communities together.

Portuguese economic literature of the second half of the 16th century exhibits innovative analyses with regard to: (*a*) an abstract conceptualization of the market as a space wherein to promote individual and public interests, and as a mechanism to reveal the value of goods exchanged; (*b*) a comparison of the advantages and disadvantages of a monopolistic organization of trading circuits; (*c*) the link between the real and the monetary spheres of the economy and an early version of a quantity theory of money; and *(d)* the doctrinal legitimization of individual gains arising from mercantile activity, for example in the case of exchange and insurance contracts. Handling such different subjects required new ways of thinking. However, in this literature, produced by merchants, theologians and court counsellors, there was no change in the theological and ethical foundations or the method on which the new elements were based. Despite the adaptations needed to interpret the new realities presented by the Discoveries, Portuguese thought continued to rest on moral and religious ideas.

During the period of the dynastic union that for 60 years (1580–1640) kept Portugal under the direct control of the Spanish crown, economic ideas started to be based on the standards of the so-called bullionist literature. However, the political restoration movement that started in 1640 initiated a search for economic strategies for consolidating independence and political sovereignty.

The tactics recommended were varied. Some favoured aiming for either balanced trade or a surplus, others favoured the introduction of monetary regulation (considering an intuitive approach to the relationship between money flows and prices), while yet others favoured the growth of the population to ensure an increase in output and tax revenues. Portuguese authors managed to receive and disseminate, almost simultaneously, different types of foreign contemporary economic literature. They adapted and used analytical constructs and economic policy proposals provided not just by Spanish but also by Italian mercantilists (particularly their proposals regarding population as a means to increase wealth), the English (the balance of trade doctrine and proposals to set up regulated companies for foreign commerce) and the French (manufacturing policy). The policy of protection for manufacturing proposed by Duarte Ribeiro de Macedo (1675), adopted at the close of the 17th century, clearly illustrates this process of assimilating ideas and economic guidelines into a national economic development strategy.

The new commercial framework imposed by the Methuen Treaty with England in 1703 did not silence supporters of this strategy. Protectionism continued to attract support, and contributed to the shaping of an entrenched tradition. During the government of Marquis of Pombal (1750–77), a protectionist economic policy was extensively applied, especially through the establishment of monopolistic companies in both commercial and productive economic activities.

Enlightened Political Economy

From the late 18th century, the development of political economy reflects the wave of economic, social, cultural and political transformations taking place throughout Europe, known as the Enlightenment. During this period, the discourse of Portuguese economists – particularly that represented by the publications of the Royal Academy of Sciences of Lisbon (Cardoso 1990–91) – reveals some familiarity with Physiocratic doctrines and principles. The primary aim of these discourses, which helped create a climate receptive to laissez-faire ideology, was the abolition of the internal barriers and excessive regulations of the *ancien régime,* which were considered as obstacles to the smooth working of the domestic market.

The dissemination of Smithian political economy was furthered by the same concerns, particularly after 1803. The ideas of both Smith and the French Physiocrats were valued as possible guidelines for a successful state-led process of social and economic change; for this reason, the reading that was made of Smith's works in Portugal by J.J. Rodrigues de Brito (1803–1805) and José da Silva Lisboa (1804) focused mainly upon the feasibility of the systems of political economy that were encouraged by François Quesnay and by the *Wealth of Nations.* Aided by Jean- Baptiste Say, Adam Smith was rapidly acknowledged as the true founder of modern political economy.

Notwithstanding the unanimous acknowledgement of the importance of Smith's economic thought in the subsequent spread of classical political economy, which led to its eventual institutionalization as a separate area of study, the English classical school did not have a significant effect on Portuguese economics. As English economic success was undeniable, and as England continued to be Portugal's principal commercial, political and military ally, the lack of a marked preference for English economics may, at first sight, seem strange. However, several factors explain it. Portugal's problematic political and diplomatic circumstances, after the first signs of the Brazilian desire for independence (1814), made it clear that England's support for free trade could be harmful. After this date, a compromise was gradually established regarding the appropriateness of the principles supported by English political economists: Acúrsio das Neves (1814–1817) made a clear distinction between the virtues of domestic liberalization and the need for prudence at the international level. On the other hand, the need to simplify and popularize political economy was seemingly better fulfilled through Continental political economy. Although some French, German and Italian works might be less vigorous analytically, they provided both an explanation and a critical assessment of many of the English school's doctrines.

Given that Continental political economists were as concerned as the Portuguese with the consequences of the English system, it was only natural that they eventually had a greater impact in Portugal than the more specific, abstract ideas of Ricardo and his followers. This is particularly evident in the first discussions regarding the choice of a political economy handbook, for no one suggested the use of an English author. The first attempts to write a Portuguese text were based either on Say's work (Manuel de Almeida) or on Storch's *Treatise* (José Ferreira Borges). In the 1840s, Forjaz de Sampaio (1841) was to write a handbook inspired by the approach developed by Karl Friderich Rau, while Marnoco e Sousa (1910), the most celebrated early 20th-century professor at the University of Coimbra, came under the spell of E.R.A. Seligman.

Establishing a Canon

The overthrow of the *ancien régime* in 1820 in a liberal revolution did not change the previous misgivings about some aspects of classical political economy, particularly those regarding international free trade.

Concern with national economic development, coupled with a suspicion that English political economy was biased in favour of English interests, led to the prevalence of an approach that was quite similar to the one that was later to be developed in Friedrich List's national system of political economy. Between 1820 and 1850, a significant number of Portuguese authors insisted that several principles of classical economic theory and policy were abstruse and therefore inappropriate for steering the development of their own country. As a result, not just the doctrine of free trade but also the theories of population, rent, diminishing returns and the stationary state were dismissed either as wrong or as being solely applicable to the more advanced English circumstances.

When, in 1836, political economy was eventually accepted as an academic discipline, and situated within legal studies, the critical stance regarding the selection of both authors and doctrines to be taught was reinforced. On the one hand, since political economy was mainly taught for the benefit of lawyers, teachers were naturally expected to emphasize the relations between economic laws and legislative action. On the other hand, since they were scholars and not pamphleteers, teachers were meant to adopt an unbiased approach to political economy, not supporting any single school of economic thought. The ensuing eclecticism had the beneficial consequence of allowing for a continuous updating process, teachers such as Adrião Forjaz de Sampaio (between the late 1830s and the early 1870s) or Marnoco e Sousa (between 1900 and 1916) always being ready to mention each and every new school of economic thought that came to their knowledge. This same concern eventually led to significant space being allotted in Portuguese law schools to the study of the history of economic thought.

When coupled with a constant awareness of the national conditions that could make some of the doctrines of political economy unworkable, and a mistrust regarding excessively abstract formulae, this eclecticism also helps us to explain why marginalism and neoclassical economics had less impact than the sociological and historical schools of the second half of 19th century.

The Rejection of Mainstream Economics

At the start of the 20th century, investment in the development of theoretical abstractions was still generally deemed by Portuguese authors not to be an essential part of their role as economists, for they thought that they should concentrate on the task of identifying and solving present economic and social problems. Therefore, even if the conceptual advances made by the marginalist revolution and the theoretical apparatus developed by neoclassical economists both in Europe and in the United States of America did have repercussions, these did not lead to any effort towards furthering economic analysis per se. Jevons was partly translated, and the doctrines of the Manchester, Austrian and Lausanne schools were closely summarized and scrutinized (either approvingly or disapprovingly). But on most occasions, these foundations of the modern canon of economic analysis were laid in Portugal in an eclectic manner, devoid of any noticeable tendency to claim that there were undisputable economic principles (see Almodovar and Cardoso 2001).

In the rare cases where this attitude did not prevail, the reaction was quite vigorous. António Horta Osório (1911) aimed to establish the importance of the mathematical method in political economy in the context of the development of the general equilibrium theory of the Lausanne School.

The Paretian distinction between utility and ophelimity was used by Osório as the foundation of his view of economic science: economics was defined as a disciplinary field restricted to the study of a small part of human behaviour, while psychology was portrayed as the global science pertaining to the study of human action. Consequently, economics would be no more than one of the branches of the general study of mankind. For him, pure economics was an abstract and experimental science which had to evaluate its scientific character, like all the exact sciences, not through the practical utility of its conclusions but chiefly by establishing the exactness of its formal internal logic.

When Osório was writing, the comprehensibility of this methodological attitude was problematic, even for those learned in economic matters, not to mention the fact that general equilibrium in exchange was far from being acknowledged by mainstream economics – something that happened only in the 1930s with the neoclassical synthesis.

In an environment that was clearly unsympathetic both to pure theory, and to any claims for the supremacy of one school, Osório was condemned to oblivion and his book was dismissed. Such an outcome symbolized the reiteration by early 20th century Portuguese economists of traditional views regarding the usefulness and role of political economy. At a time when economics was moving decisively towards establishing itself as a science, Portuguese authors stood aloof from that process. At a time when Portugal's political situation was characterized by a fair degree of cosmopolitanism, economic discourse remained focused on its own political implications, musing on topics of an ideological nature.

From Corporatism to Keynesianism

The traditional Portuguese unwillingness to abide by any single school of economic thought faded away only when faced with the ambitious state-driven project of building up a new type of political economy, that of the corporatist state (see Almodovar and Cardoso 2005).

The political and economic experiment of corporatism represents one of the most interesting stages in the study of the historical evolution of economic thought in Portugal. The corporatist movement in itself is part of the broader movement of authoritarian experiments that took place on the European and international political scene, especially after the First World War. The restlessness of many social and political sectors, which were discontented with the performance of liberal and socialist regimes, paved the way for a search for an alternative to both capitalism and socialism.

Corporatism was therefore offered as a third way between existing regimes, and its supporters claimed that it provided the sole reliable answer to the ongoing social, political and economic turmoil. Portuguese authors like Pires Cardoso and Marcelo Caetano joined their French and Italian counterparts in the effort to develop an economic point of view that could match the ethical and philosophical base provided by the philosophy of Thomas Aquinas. The final outcome was an economic doctrine openly against utilitarianism and in favour of the gradual establishment of a new type of economic agent – the so-called *homo corporativus* – that would be capable of re-embedding social values and aims into its own scale of preferences. This quest did not stop the reception of Keynes, and particularly those ideas that could be seen as a critique of the idea of a self-regulated market and as an appeal to some state intervention in favour of a more socialized economy. Keynes was fairly well-known in Portugal from the mid-1920s onwards. However, the reception and assimilation of the *General Theory* occurred more than ten years after its publication, when Fernando Pinto Loureiro (1948) and Luis Simões de Abreu (1949) published the first extended reviews and digests of J.M. Keynes's major work. Therefore, it was only after 1950, under the impulse provided by the works of António Pinto Barbosa (1950), Jacinto Nunes (1956) and F. Pereira de Moura (1964), that Keynesian concepts began to play a significant role in newly established courses on macroeconomics, public finance, development economics and econometrics.

Despite this process, the reception given to the propositions of the *General Theory* at the political level was ambiguous and much less enthusiastic. One of the most notable aspects of the economic strategy followed in Portugal throughout the 1930s and 1940s was the enactment of legislation to set up industrial companies and to control industrial activity. The basic aim was not just to prevent Portuguese industry from being disturbed by internal or external competition but to keep in check and organize industrial growth and the overall process of economic development. In such a context, where the pace of development was restricted and a balanced budget and the preservation of the country's gold reserves were praised, Keynesian economics could hardly figure prominently.

This vision of economic life was shared by the various political assemblies and executive bodies responsible for directing economic and financial policy. As a result, whenever they incorporated Keynesian ideas, they did so in a watered-down and superficial manner. In fact, it can be said that in the post-war period no type of short- run macroeconomic policy was ever developed: the factors which would normally justify it – such as unemployment and external disequilibrium – did not represent real problems for the Portuguese economy. Something similar occurred in relation to long-run economic policy. The first five-year development plan (1953) was totally insensitive to the assessment of its impact in macroeconomic terms. The second plan in 1958 contained, in the explanation given for its design, projections based on a Harrod–Domar growth model, but this was no more than a rhetorical device.

Throughout the period that we have been considering here, Portuguese economic policy remained faithful to corporatist principles, coupled with a traditional model of empirical, and essentially descriptive, economic studies, without any visible influence of Keynesian concepts.

Concluding Remarks

The first Portuguese university institution created specifically for the teaching of economics was formally founded in 1933. A profound reform of its curricular and pedagogical structure took place in 1949, involving the replacement of essentially technical courses in the fields of commerce, bookkeeping, accounting, customs and diplomatic services with more general courses in economics and finance. Only after that did the full incorporation of a neoclassical approach begin to take place, in the form of a synthesis with Keynesian thought. However, integration into the international mainstream was held back by the resilience of the traditional Portuguese attitude regarding economic knowledge, which was to try to take over the doctrinal and political ingredients that best fitted the search for a specifically Portuguese route to economic development. At all of the most significant moments in the evolution of economic thought in Portugal, we find this attempt to select and adapt existing economic ideas to Portuguese circumstances. Inquisitiveness regarding alternative routes to economic progress, coupled with a pragmatic view regarding economic policy guidelines, favoured a continuous oscillation between schools of economic thought and the emergence of eclecticism. As a consequence, Portuguese economic thinking retained its links with law, ethics, politics, and sociology; and it took a long time to accept the autonomy and analytical competence of economics (see Almodovar and Cardoso 1998).

A further example of the tendency to eclecticism prevalent among Portuguese economists was the impact of the structuralist and developmentalist economic ideas in the 1960s and in the 1970s, through the influence of the Latin American economists concerned with the problems of underdevelopment and with the political responses to overcome it. However, this influence was superseded by the process of harmonization resulting from the democratic revolution of 1974 and Portugal's integration into the European economy in 1986. The considerable institutional changes that then occurred have largely contributed to the smooth reception and institutionalization of both macroeconomic and microeconomic principles and applications. As a result of this integration process, eclecticism has gradually given way to approaches that conform much more closely with the international mainstream.

See Also

- ▶ France, Economics in (Before 1870)
- ▶ France, Economics in (After 1870)
- ▶ Heterodox Economics
- ▶ Historical School, German
- ▶ Italy, Economics in
- ▶ Keynesianism
- ▶ Mercantilism
- ▶ Physiocracy
- ▶ Spain, Economics in

Bibliography

Abreu, L.S. 1949. Algumas notas sobre a economia de Keynes [A few notes on the economics of Keynes]. *Revista de Economia* 1(3): 15–25.

Almodovar, A., and J.L. Cardoso. 1998. *A history of Portuguese economic thought*. London/New York: Routledge.

Almodovar, A., and J.L. Cardoso. 2001. From learned societies to professional associations: The establishment of the economist profession in Portugal. In *The spread of political economy and the professionalisation of economists*, ed. M. Augello and M. Guido. London/New York: Routledge.

Almodovar, A. and J.L. Cardoso. 2005. Corporatism and the economic role of government. In *The role of government in the history of economic thought*, ed. S. Medema and P. Boettke. Annual Supplement to *History of political economy*, vol. 37. Durham/London: Duke University Press.

Barbosa, A.M.P. 1950. *Economia*. Lisboa: ISCEF.

Brito, J.J.R. 1803–1805. In *Memórias Políticas sobre as Verdadeiras Bases da Grandeza das Nações e Principalmente de Portugal* [Political Memoirs on the True Basis of the Greatness of Nations, Especially of Portugal], ed. J.E. Pereira. Lisboa: Banco de Portugal, 1992.

Cardoso, J.L., ed. 1990–91. *Memórias Económicas da Real Academia das Ciências de Lisboa, para o Adiantamento da Agricultura, das Artes, e da Indústria em Portugal, e suas Conquistas, 1789–1815* [Economic Transactions of the Lisbon Academy of Sciences for the Advancement of Agriculture, Industry and Commerce in Portugal and its Domains], 5 vols. Lisboa: Banco de Portugal.

Lisboa, J.S. 1804. Princípios de Economia Política. In *Escritos Económicos Escolhidos (1804–1820)* [Selected Economic Writings], vol. 1, ed. A. Almodovar. Lisboa: Banco de Portugal, 1993.

Loureiro, F.P. 1948. *Sobre a Introdução ao Estudo da Nova Economia Keynesiana* [Introduction to the Study of the New Keynesian Economics]. Coimbra: Coimbra Editora.

Macedo, D.R. 1675. Discurso sobre a Introdução das Artes no Reino [Discourse on the Introduction of Manufactures in the Kingdom]. In *Antologia dos Economistas Portugueses – Século XVII* [Anthology of Portuguese Economists – 17th Century], ed. A. Sérgio. Lisboa: Biblioteca Nacional, 1924.

Moura, F.P. 1964. *Lições de Economia* [Lectures in Economics]. Lisboa: Clássica Editora.

Neves, J.A. 1814–1817. *Variedades sobre Objectos Relativos às Artes, Comércio e Manufacturas, Consideradas Segundo os Princípios da Economia Política* [Varieties on Subjects Concerning Industry and Commerce, Considered Under the Principles of Political Economy], ed. A. Almodovar and A. Castro. Porto: Afrontamento, 1994.

Nunes, M.J. 1956. *Rendimento Nacional e Equilíbrio Orçamental* [National Income and Budgetary Equilibrium]. Lisboa: Editorial Império.

Osório, A.H. 1911. *A Matemática na Economia Pura: a Troca* [Mathematics in Pure Economics: Exchange], ed. M. Farto. Lisboa: Banco Portugal, 1996.

Sampaio, A.F. 1841. Elementos de Economia Política e Estadística [Elements of Political Economy and Statistics]. In *Estudos e Elementos de Economia Política (1839–1874)* [Studies and Elements of Political Economy], vol. 2, ed. A. Pedrosa. Lisboa: Banco de Portugal, 1995.

Sousa, J.M. 1910. *Ciência Económica* [Economics], ed. M. de Fátima Brandão. Lisboa: Banco de Portugal, 1997.

Positive Economics

Richard G. Lipsey

Abstract

'Positive economics' refers to the view that economic theories consistent with all conceivable observations are empirically empty and that empirically useful theories need to be consistent with existing observations (thus passing the 'sunrise test') and predict something new. It is neither logical positivist, nor operationalist, nor naïve falsificationist; nor is it based on strict dichotomies between positive and normative statements and between positive analysis and normative advice. It rejects the views that theories can assist understanding the world without making refutable statements about it; that theories can be criticized only on their own terms; and that all distinctions inhibit useful discourse.

Keywords

Agassi, J.; Archibald, G.; Assumptions; Barriers to entry; Bekar, C.; Blaug, M.; Caldwell, B.; Carlaw, K.; De Marchi, N.; Economic history; Falsificationism; Friedman, M.; Game theory; Hahn, F.; Hodgson, G.; Hutchison, T.; Industrial organization; Klappholz, K.; Labour economics; Lind, H.; Lipsey, R. G.; Logical positivism; Methodology; Monopolistic competition; Operationalism; Policy advice; Popper, K.; Positive economics; Positive–normative distinction; Robbins, L.; Samuelson, P.; Scale effects; Sunrise test; Testing; Theory appraisal; Value judgements; Verificationism; Wong, S.

JEL Classifications

B4

The term 'positive economics' refers to some specific views about what makes economics a science. According to some of the most influential 19th century English economists, positive economics consisted of propositions or 'laws' concerning real-world events that were derived from intuitively self-evident assumptions. Facts were to be used, therefore, as illustrations of theories, not as tests. To give policy advice, the propositions of positive economics had to be combined with value judgements. An elegant 20th century statement of this view of 'scientific economics' was given by Robbins (1935). Not surprisingly, these economists were, as Blaug (1992) has argued, 'verificationists' who shielded their theories from empirical refutation.

The term was used by such 20th century writers as Friedman (1953) and Lipsey (1963) to refer to what they regarded as scientific economics: non-normative theories whose assumptions were not necessarily self-evident and whose implications were to be judged by empirical observations. Karl Popper provided the methodological underpinning of these works, underpinnings that were either implicit (as in Friedman's case) or explicit (as with most other writers). Terence Hutchison (1938) was the first to introduce Popper's ideas to economists, although he did not describe his work as positive economics. Blaug (1992) and Hutchison (1992) provide excellent formulations of the main tenets of modern positive economics, along with criticisms of both its main detractors and advocates of other views.

Friedman (1953, pp. 7–8) stated the sense in which he understood the term 'positive' when he wrote: 'The ultimate goal of positive science is the development of a "theory" or "hypothesis" that yields valid and meaningful (i.e., not truistic) predictions about phenomena not yet observed. . . .

only factual evidence can show whether it is . . . tentatively "accepted" as valid or "rejected" '. Shortly after my textbook appeared, I wrote:

> I tried to break away from the treatment of economic theory as revealed truth and to emphasize from the outset the very tentative nature of much of our economics . . . to say . . . to the student 'you cannot have both certainty and empirical relevance'.. . . The adjective 'positive' in the title of the book was [partly an allusion to the positive–-normative distinction and] partly an allusion to the distinction between positive (i.e., empirical) Versus a priori methods of judging between theories. (Lipsey 1964, pp. 370–1)

To this end, the concluding chapters in several parts of my textbook discussed 'measurement', 'tests', and 'criticisms' of the theories already presented.

Some Criticisms and Misunderstandings

The history of positive economics at the London School of Economics' Staff Seminar on Methodology, Measurement, and Testing (M^2T seminar) has been well described by de Marchi (1988) – although I disagree with most of his conclusions on pages 162–3. Ours was a crusade for making economic theories empirically relevant and for rejecting intuition as the test of validity, replacing it with empirical testing. Positive economics, as we conceived it, had two main messages. First, if an economic theory is to be about the real world, it must be possible to imagine observations that would be conflict with it. If conflicting observations cannot even be imagined, the theory is compatible with all states of the world and hence empirically empty. A great advance in making theory more relevant would be achieved if today's editors insisted that each author state what factual observations would conflict with his or her theory, and, if there were none, to state the theory's purpose. Second, a new theory should be compatible with ('explain') some existing facts and suggest some new one(s).

We were subject to misunderstandings, as well as to criticisms, from those who disagreed with our main message. Here are some of the most influential.

1. Its philosophical base was thought by many critics to be logical positivism, which we rejected in favour of Popper's methodology.
2. Samuelson (1948) was partly responsible for another confusion when he stated a similar view on testability but espoused a version of operationalism. We never took an operationalist position, arguing only that *all* those parts of a theory that did say something empirically should be open to empirical testing. Subsequently, Wong (1978) argued – correctly in our view – that, since even such simple 'entities' as prices and quantities are theoretical concepts that do not exactly correspond to real-world entities, theories cannot be expressed solely in operational terms.
3. Friedman set off a long debate on the testability of assumptions that helped to discredit positive economics to many. We disagreed with Friedman, arguing that, if empirically correct predictions were deduced from a set of empirically false assumptions, this called for further serious study, not complacency. See Blaug (1992, pp. 91–7) for a full discussion.
4. Contrary to the tenets of positive economics, Friedman used his essay on methodology to dismiss monopolistic competition as adding nothing that could not be learned from a judicious combination of perfect competition and monopoly. That this was not our position was shown when this, and similar arguments of others in the Chicago school, were criticized by members of the M^2T seminar (Klappholz and Agassi 1959; Archibald 1961).
5. We were accused of being naive falsificationists. Although we may have been at the outset, we soon refined our position as a result of experience and accepted that 'we cannot get a categorical disproof of an hypothesis' (Lipsey 1975, p. 46), which statement was followed by a long passage on what can be learned from apparent refutations. Another member of the M^2T seminar, Archibald (1967), argued that empirical testing could at best establish the balance of probabilities between two conflicting theories rather than refuting either categorically. We did not, however, as implied by de Marchi (1988, p. 162)

P

give up on positive economics just because we abandoned naive falsification; indeed, many of us went on to do significant empirical work.

6. Some critics argued that positive economics was merely what Mark Blaug calls 'conformatism', asking only that a theory be consistent with known facts. From the very outset we accepted Popper's criticism that theories that explained only already known facts were being subjected to a 'sunrise test' from which we could only learn that the theorist was ingenious enough to build a theory that jumped through predetermined hoops.
7. Others argued that we naively accepted the earlier economists' strict dichotomy between positive and normative statements. We quickly discarded this view. Lipsey (1963, p. 4, n. 1) introduces a discussion of this matter thus: 'Philosopher friends have persuaded me that when pushed to its limits, the distinction between positive and normative becomes blurred or breaks down completely.' However, the blurring did not stop us from arguing that the ability to distinguish what one thinks is true from what one would like to be true is critical to all science.
8. In a similar but not identical vein, yet others argued that we naively accepted the strict division that the earlier economists had made between positive economic analysis and normative advice. My first exposure to policy advising in 1962 disabused me of that idea. As I later put it: 'The economic adviser and the policy-maker are involved in a complex human relationship, entangled in various uncertainties and communicating with each other through an inevitable haze of emotional reactions. Economists may strive towards an ideal of communicating their knowledge as objectively as possible, but objectivity remains an ideal that guides their actions, not a reality that fully describes them' (Lipsey 1981, p. 35). For detailed discussions of the history of the distinction.

We rejected many other methodological approaches that were either implicit or explicit criticisms of positive economics, of which the following are examples. First was the view of many pure theorists, such as Hahn (1984, pp. 44–5), criticized by Blaug (1992, pp. 164–5) and Hutchison (1992, p. 43), that theories can somehow add to our understanding of the world without making testable statements about it. Second, there was the view subsequently articulated by Caldwell (1982) that falsification is too strong and that we can criticize each school of thought only on its own terms. As Hutchison pointed out, this amounts to rejecting in principle any method of discriminating between alternative theories. Third, in reaction to the view that all distinctions inhibit full discourse, we maintained that distinctions help to structure arguments and, without them, there is anarchy of discourse.

Positive Economics Today

What is the fate of positive economics today? While many economists pay lip service to the view that economic theories should make testable predictions about the world and that the ultimate arbiter of different theories is empirical evidence, many research programmes do not show this as their revealed preference. Theoretical articles that do no more than state and pass sunrise tests abound.

The modern version of industrial organization has had most of its empirical content eliminated. Students who used to learn institutional material about such 'practical' matters as competition policy, and who studied empirical information about scale effects and entry barriers, often today know little more than game theory. (See Lipsey 2001, for full discussion.)

The new formalism has given rise to the belief among some theorists (although not typically among those who do game theory) that the more general a theory is, the better it must be. But this assumption ignores the fact that the more general a theory is, the less empirical content it is likely to have since, by ignoring the specific context in which many problems arise, it becomes impossible to analyse them in depth. (See Hodgson 2001, for a full discussion.) One set of examples, criticized at length in Lipsey et al. (2005, pp. 466–7), is found in those modern growth theories that use an aggregate production function devoid of institutions or anything that distinguishes economies with various degrees of development.

Not a few theories are devoted to explaining mere possibilities. Typically someone develops a simple ‘Mark I model’ on some matter such as the effects of rent controls and draws strong policy conclusions from it; someone else comes along with a Mark II model saying ‘if I add some not implausible complexity, the model’s predictions and policy conclusions are altered’. Then someone does the same to the Mark II model, and so on. (For a case study see Lind 2007.) Although it is possible to learn something from all exercises, this sort of research programme tells us little more than that very simple and more complex theories on the same issue do not usually have identical predictions.

Many research programmes are ‘internally driven’, by which I mean that they are driven by their own internal logic. Investigators seek to understand problems created by the models that they are using rather than deriving their problems from observations. In contrast, an ‘externally driven research program’ (EDRP) is one that is driven and constrained by observed facts. A perusal of the literature will show IDRPs to be at least as common as EDRPs. (For examples, see Lipsey 2001.)

On a personal level, the revolution that I tried to create in textbook writing through *An Introduction to Positive Economics* has slowly dwindled – in spite of its being the dominant text book in the UK for decades, being widely used throughout the Commonwealth, and having significant sales in its US adaptation, initially co-authored with Peter Steiner. As a result of constant criticism from teachers who wanted to present only mainline economics, the criticism and testing chapters were slowly eroded – much faster in the US editions than in the UK ones. Today, all too many modern theory textbooks at all levels, from basic to advanced, present current economic theories as if they were revealed truth, paying little attention to controversies and alternative theories.

Finally, I ask what the real successes of positive economics are. As already mentioned, most economists pay a least lip service to the ideal that economic theories are meant to tell us something about the real world through potentially testable hypotheses. The journals are full of empirical observations, many of which are extremely useful – although many others are used in the non-informative types of theory mentioned earlier. In some empirically oriented fields, such as economic history and labour economics, the ideal of positive economics does come close to realization. For example, much work in labour economics seeks to establish empirical relations, such as those between the characteristics of a person’s schooling and his or her lifetime earnings. It is a matter of taste whether one interprets these studies as hypothesis testing or just establishing statistical relations, but either way they are important.

So the ideas of positive economics are still present – more strongly in some fields than in others – though many economists reject them, as shown either explicitly by their methodological pronouncements or implicitly by their research practices.

See Also

- ▶ Economics, Definition of
- ▶ Friedman, Milton (1912–2006)
- ▶ Methodology of Economics
- ▶ Positivism
- ▶ Robbins, Lionel Charles (1898–1984)
- ▶ Value Judgements

Bibliography

Archibald, G. 1961. Chamberlain versus Chicago. *Review of Economic Studies* 29: 316–327.

Archibald, G. 1967. Refutation or comparison? *British Journal for the Philosophy of Science* 17: 279–296.

Blaug, M. 1992. *The methodology of economics*. Cambridge: Cambridge University Press.

Blaug, M. 1998. The positive–normative distinction. In *The handbook of economic methodology*, ed. J. Davis, D. Hands, and U. Maki. Cheltenham: Edward Elgar.

Caldwell, B. 1982. *Beyond positivism: Economic methodology in the twentieth century.* London: Allen and Unwin.

de Marchi, N. 1988. Popper and the LSE economists. In *The popperian legacy in economics*, ed. N. de Marchi. Cambridge: Cambridge University Press.

Friedman, M. 1953. The methodology of positive economics. In *Essays in positive economics*. Chicago: University of Chicago Press.

Hahn, F. 1984. *Equilibrium and macroeconomics*. Oxford: Basil Blackwell.

Hodgson, G. 2001. *How economists forgot history: The problem of historical specificity in social science*. London: Routledge.
Hutchison, T. 1938. *The significance and basic postulates of economic theory*. London: Macmillan.
Hutchison, T. 1992. *Changing aims in economics*. Oxford: Blackwell.
Klappholz, K., and J. Agassi. 1959. Methodological prescriptions in economics. *Economica* 26: 60–74.
Lind, H. 2007. The story and the model done: An evaluation of mathematical models of rent control. *Regional Science and Urban Economics* 37: 183–198.
Lipsey, R.G. 1963. *An introduction to positive economics*, 1st edn. London: Weidenfeld and Nicolson.
Lipsey, R.G. 1964. Positive economics in relation to some current trends. *Journal of the Economics Association* 5: 365–371. Reprinted in Lipsey (1997).
Lipsey, R.G. 1975. *An introduction to positive economics*, 4th edn. London: Weidenfeld and Nicolson. The page referenced is reprinted in Lipsey (1997).
Lipsey, R.G. 1981. Economists, policy makers and economic policy. In *Economic policy making in Canada*, ed. D. Smith. Montreal: C.D. Howe Institute. Reprinted in Lipsey (1997).
Lipsey, R.G. 1997. *The selected essays of Richard Lipsey. Volume I: Microeconomics, growth and political economy*. Cheltenham: Edward Elgar.
Lipsey, R.G. 2001. Successes and failures in the transformation of economics. *Journal of Economic Methodology* 8: 169–201.
Lipsey, R.G., K. Carlaw, and C. Bekar. 2005. *Economic transformations: General purpose technologies and long-term economic growth*. Oxford: Oxford University Press.
Robbins, L. 1935. *An essay on the nature and significance of economic science*, 2nd edn. London: Macmillan.
Samuelson, P. 1948. *Foundations of economic analysis*. Cambridge, MA: Harvard University Press.
Wong, S. 1978. *The foundations of Paul Samuelson's revealed preference theory: A study by the method of rational reconstruction*. London: Routledge and Kegan Paul.

Positivism

Bruce Caldwell

Abstract

The article identifies the major tenets of logical positivism and its successor, logical empiricism, two important movements within 20th-century philosophy of science. It then documents some of the arguments that led to the decline of positivism in the latter half of the 20th century. The impact of positivist ideas on the work of economists writing about economic methodology is examined in a final section.

Keywords

Blaug, M.; Braithwaite, R.; Carnap, R.; Comte, A.; Confirmationism; Falsifiability; Friedman, M.; General equilibrium; Hahn, H.; Hempel, C.; Hume, D.; Hutchison, T.; Hypothetico-deductive model; Induction; Logical empiricism; Logical positivism; Mach, E.; Machlup, F.; Methodology of economics; Mill, J. S.; Mises, L. von; Nagel, E.; Neurath, O.; Pareto, V.; Philosophy of science; Popper, K.; Positivism; Robbins, L.; Samuelson, P. theory appraisal; Verifiability; Vienna circle

JEL Classifications

B1

Positivism and the Philosophy of Science

The term 'positivism' was coined in the second quarter of the nineteenth century by one of the founders of sociology, Auguste Comte. Comte believed that human reasoning passes through three distinct historical stages: the theological, the metaphysical, and the scientific. In the theological stage, natural and social phenomena are explained by reference to spiritual forces. In the metaphysical stage, 'ultimate causes' are sought to explain such phenomena. In the scientific stage, attempts to explain phenomena are abandoned, and scientists seek instead to discover correlations among phenomena (Comte 1830–1842). Another important figure in the development of *classical positivism* was the physicist Ernst Mach (1886), who propounded a 'fictionalist' view of theories. Scientific theories are useful mnemonic devices, but progress in science occurs only when such useful fictions are replaced by statements which contain only observation terms. Though both Comte and Mach had some influence on the writings of economists (Comte influenced J.S. Mill and Pareto, Mach was mentioned in passing by Samuelson and Machlup), their primary influence was on the ideas of certain twentieth century philosophers of science, the logical positivists.

Logical Positivism

The major tenets of *logical positivism* were developed in the 1920s by Moritz Schlick, Herbert Feigl, Kurt Gödel, Hans Hahn, Otto Neurath, Friedrich Waismann, Rudolf Carnap and other members of the famous Vienna Circle. Logical positivism was a radically empiricist philosophical position, and its founders believed it marked a new beginning for philosophical inquiry. The goal of all philosophical analysis was henceforth to be the logical analysis of the knowledge claims of the positive, or empirical, sciences: hence the label 'logical positivism'.

The first task facing the logical positivists was to define what constitutes a knowledge claim. Their solution was to analyse the logical form of statements. Only statements that are either analytic (such as definitions) or synthetic (testable statements of fact) qualify as cognitively significant, or meaningful. All other statements lack cognitive significance: they are meaningless, metaphysical, non-scientific. Analyses that make use of such statements may express emotional stances, or 'general attitudes towards life', or moral valuations, but they do *not* express knowledge claims.

To put their programme into operation, the logical positivists needed an objective criterion of cognitive significance which could be used to distinguish synthetic statements from meaningless ones. One early solution was the principle of verifiability: a synthetic statement has meaning only if it is verifiable. Unfortunately, statements of universal form (for example, 'all ravens are black'), which are frequently encountered in science, are unverifiable. Other criteria included falsifiability, Ayer's weak verifiability, Carnap's translatability into an empiricist language, and confirmability. None of these was able to resolve the problem conclusively, however. Another dilemma was posed by the presence of theoretical terms in statements made by scientists. Some positivists followed Mach in insisting that they should be eliminated from science, while others argued that such statements should be retained. A final element of the logical positivist programme was an emphasis on the unity of science, variously defined as meaning that all true sciences share a common method, that the results of all sciences should ultimately be expressible in a common physicalist language, or that the results of the various sciences should be integrated, better to assist the scientific planning of society.

Logical Empiricism

Hahn died in 1934, and Schlick was murdered in 1936 by an insane student. But it was Hitler's rise to power, and the subsequent flight of intellectuals, that primarily caused the disintegration of the Vienna Circle in the 1930s. Logical positivism was modified and ultimately replaced over the next two decades by a more analytically austere form of positivist thought, *logical empiricism.* Though differences exist in their analyses, philosophers who have contributed to this later tradition include Carnap, Ernest Nagel, Carl Hempel and Richard Braithwaite.

There were six major tenets of the logical empiricist programme. First, the *unity of science thesis* was narrowed to mean only a unity of scientific methods. The next three had to do with the structure and appraisal of theories. The *hypothetico-deductive model of theory structure* states that all sciences employ theories, which may be represented formally as axiomatic, hypothetico-deductive structures. Such structures have no empirical import until some of their elements (usually the deduced theorems, or predictions of the theories) are given an empirical interpretation via the use of correspondence rules. Not every statement will have an empirical interpretation. Those containing theoretical terms, in particular, will not be interpretable. Are such sentences then meaningless? Not at all; according to the *indirect testability thesis* such sentences gain cognitive significance indirectly when the theories in which they are embedded are confirmed. Finally, concerning the questions of demarcation and theory assessment, logical empiricists settled on *confirmationism* as their primary criterion of theory appraisal. A theory is scientific if it is testable; test instances confirm or disconfirm the theory; the acceptability of the theory depends on its degree of confirmation. Degree of confirmation is measured by such things as the quantity and precision of favourable test outcomes, the precision of procedures of observation and measurement, the variety of supporting evidence, and whether new test situations support the hypothesis. Additional non-empirical criteria of appraisal (for example, simplicity, elegance, fruitfulness, generality, extensibility) may also be invoked if theory choice on empirical grounds yields no preferred theory. The last two tenets of logical empiricism concerned the logic of scientific explanation. All explanations in science must be expressible in the form of a deductive argument in which an explanandum, a sentence describing the event to be explained, is logically deduced from an explanans. The explanans contains a group of sentences, some of which express initial conditions, and at least one of which states either a general or a statistical law. The *deductive-nomological* and *inductive-probabilistic covering law models of scientific explanation* take their names, then, from the types of laws (general or statistical) used in the explanations. Additionally, logical empiricists believed in the *symmetry thesis*: explanation and prediction are structurally symmetrical, the only difference between them being one of temporality. In the case of an explanation the phenomenon described in the explanandum has already taken place, whereas in the case of a prediction it has not yet occurred.

As documented in Suppe (1977), logical empiricist ideas (sometimes dubbed 'the received view') came under heavy attack in the mid-twentieth century. The viability of both the hypothetico-deductive model of theory structure and the indirect testability thesis depended on one's ability to draw a clear distinction between observational terms (terms that refer to observables, to 'brute, atomic facts') and non-observational, theoretical terms. Unfortunately, in many sciences there are degrees of observability, and no hard division can be drawn between theoretical terms that refer to non-observables and non-theoretical terms that refer to observables. Furthermore, because observation itself is not a neutral activity but requires both data selection and interpretation, it was argued (by critics like Karl Popper and Norwood Hanson) that all observation is theory-dependent. Regarding confirmationism, the failure to solve Hume's problem of induction and a number of paradoxes of confirmation undercut attempts to construct an inductive logic of confirmation. In addition, Popper (1959) challenged the desirability of making statements that have a high inductive probability. Finally, many explanations in a variety of sciences could not be reconciled with the two covering law models of scientific explanation.

The Naturalistic Turn

The influence of positivism within the philosophy of science declined considerably through the 1960s and 1970s. As noted by Hands (2001), its apparent successor has been dubbed the *naturalistic turn,* an approach that, rather than laying out a priori criteria for identifying appropriate scientific practice, instead employs the tools of the sciences themselves to investigate scientific practice. There are, of course, many different scientific disciplines from which to draw such tools; some that have been used are cognitive psychology, evolutionary biology, sociology, and economics. Depending on which scientific practice is analysed, reflexivity issues may appear (for example, in using economic analysis to explain the development of economics and the behaviour of economists). Other important issues facing the naturalistic turn are choosing among the various tools on offer, and deciding whether the ensuing analysis has prescriptive implications in addition to descriptive merits. Another movement that has had less impact in philosophy of science proper, but great influence in a number of sciences including economics, derives from the work of Karl Popper. A critic of inductivism and confirmationism, the father of falsifiability and of critical rationalism, Popper had sufficient insight, foresight and longevity to influence a number of generations of philosophers of science, among them J. Agassi, W.W. Bartley III, P.K. Feyerabend and Imre Lakatos. Within economics, the work of T.W. Hutchison (for example, 1997), Mark Blaug (1992) and Lawrence Boland (2003) most directly reflect Popper's influence, while that of Wade Hands (1993) and Bruce Caldwell (1991) reflect a critical reappraisal.

In the 1990s an historical dehomogenization of the writings of the logical positivists of the Vienna Circle began. A rehabilitation of Otto Neurath, whose anti-foundationalism, advocacy of pluralism, and emphasis on scientific practice led many to see him as a precursor of the naturalistic turn, was the most notable result (Uebel 1991). Some historians and philosophers also praised his willingness to advocate the scientific planning of society and of science, to employ the philosophy of science as a tool in the restructuring of society. For these interpreters, the emergence of a more austere logical empiricism in the 1950s represented not a scientific advance but a retreat to more neutral formalism in response to the ideological pressures of McCarthyism and the cold war (for example, Reisch 2005). This interpretation parallels Philip Mirowski's (2002) historical account of the development of formalism in economics during the same period.

Positivism and Economics

There are various ways to describe the influence of positivist thought in economics.

If one focuses on the period in which positivist philosophy of science was invoked by economists, the positivist epoch spanned roughly 40 years, from the late 1930s to the late 1970s. This is not to say that during this period economists self-consciously adopted the philosophical positions outlined above. As shown in Caldwell (1994), what in fact occurred was that certain economists writing about methodology borrowed, usually somewhat haphazardly, from the language of positivism, while others invoked various positivist positions to defend or to criticize theories and practices in economics.

Four economists from this period whose writings most reflect the influence of positivism are T.W. Hutchison, Fritz Machlup, Paul Samuelson, and Milton Friedman. In the 1938 book, *The Significance and Basic Postulates of Economic Theory*, Hutchison launched an empiricist attack on the pure logic of choice, a doctrine that had been espoused and defended by Lionel Robbins 6 years earlier in his *The Nature and Significance of Economic Science* (1932). For more than 50 years, Hutchison was to continue to criticize all forms of economics that were based on untestable foundations, his targets ranging from the apriorism of Ludwig von Mises to the elaborate mathematical models of general equilibrium theory. Fritz Machlup offered one response to Hutchison with his 1955 paper, 'On the problem of verification in economics', where he invoked the indirect testability thesis to defend the use of theoretical constructs in economics against what he dubbed Hutchison's 'ultra-empiricism.' In the Introduction of his *Foundations of Economic Analysis,* Paul Samuelson (1947) borrowed from the work of physicist Percy Bridgman when he insisted that economists search for operationally meaningful theorems. The intent of Samuelson's revealed preference approach to demand theory was to place consumer theory on an observational basis. Finally, Milton Friedman's influential 1953 piece 'The methodology of positive economics' contained the famous argument that the realism of the assumptions of a theory is irrelevant; what counts in the assessment of a theory is its relative predictive adequacy and its simplicity. Though Friedman's unique brand of instrumentalist methodology owes more to the American pragmatists than to positivism, his approach came to be viewed as synonymous with positivism through the 1950s and 1960s.

Though economists today rarely invoke positivist philosophy of science in defending their preferred practices, there is plentiful evidence of its continued influence, mostly in terms of what is considered to be 'appropriate' or 'legitimate' practice, with 'positivist' often being equated with 'truly scientific'. Thus, important areas like game theory and transactions cost analysis initially encountered substantial opposition from mainstream economists because such analyses, though rich in terms of explaining diverse economic phenomena, often did not produce the sort of testable hypotheses demanded by positivist doctrine. (Strangely, during its period of dominance, general equilibrium theory was much less affected by such critiques.) Similarly, the positivist belief in the cumulative development of science tends to render less important both heterodox approaches to the discipline and the study of doctrinal history. Finally, the insistence on defining progress in terms of 'the discovery of law-like relationships' or 'better predictive ability' has fuelled a sustained growth in data collection and in computing power, the development of new econometric techniques, and a staggering increase in empirical studies. That all this has resulted in at best meagre progress (see Backhouse 1997) in establishing robust economic 'laws' and in improving forecasting power has typically engendered not a reassessment of the goals but a redoubling of resources committed to reaching them, with the attendant opportunity costs. It will be interesting to see what the entry on 'positivism' in the third edition of *The New Palgrave* reveals about its legacy in economics.

See Also

- ▶ Methodology of Economics
- ▶ Philosophy and Economics
- ▶ Theory Appraisal

Bibliography

Backhouse, R. 1997. *Truth and progress in economic knowledge*. Cheltenham: Edward Elgar.

Blaug, M. 1992. *The methodology of economics: or how economists explain*. 2nd ed. Cambridge: Cambridge University Press.

Boland, L. 2003. *The foundations of economic method: A Popperian perspective*. London: Routledge.

Caldwell, B. 1991. Clarifying Popper. *Journal of Economic Literature* 29: 1–33.

Caldwell, B. 1994. *Beyond positivism: Economic methodology in the twentieth century*. London: Routledge.

Comte, A. 1830–1842. *Cours de philosophie positive*. Paris: Bachelier.

Friedman, M. 1953. The methodology of positive economics. In *Essays in positive economics*, ed. M. Friedman. Chicago: University of Chicago Press.

Hands, D. 1993. *Testing, rationality and progress: Essays on the Popperian tradition in economic methodology*. Lanham: Rowman and Littlefield.

Hands, D. 2001. *Reflection without rules: Economic methodology and contemporary science theory*. Cambridge: Cambridge University Press.

Hutchison, T. 1938. *The significance and basic postulates of economic theory*. Reprinted, with a new preface, New York: Kelley, 1960.

Hutchison, T.W. 1977. *Knowledge and ignorance in economics*. Chicago: University of Chicago Press.

Mach, E. 1886. The analysis of sensations, ed. S. Waterlow, trans. C. Williams. New York: Dover, 1959.

Machlup, F. 1955. The problem of verification in economics. *Southern Economic Journal* 22: 1–21.

Mirowski, P. 2002. *Machine dreams: Economics becomes a cyborg science*. Cambridge: Cambridge University Press.

Popper, K. 1959. *The logic of scientific discovery*. New York: Basic Books.

Reisch, G. 2005. *How the cold war transformed philosophy of science: To the icy slopes of logic*. Cambridge: Cambridge University Press.

Robbins, L. 1932. *An essay on the nature and significance of economic science*. 2nd ed. London: Macmillan. 1935.

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

Suppe, F., ed. 1977. *The structure of scientific theories*. 2nd ed. Urbana: University of Illinois Press.

Uebel, T. 1991. *Rediscovering the forgotten Vienna Circle: Austrian studies on Otto Neurath and the Vienna Circle*. Dordrecht: Kluwer.

Post Keynesian Economics

J. E. King

Abstract

Post Keynesian economics is a dissident school in macroeconomics based on a particular interpretation of Keynes. A brief intellectual history of Post Keynesian ideas is provided, along with a discussion of some important methodological questions. Three short-period macro models are outlined: Paul Davidson's aggregate supply–aggregate demand model, Michal Kalecki's two-class model, and Hyman Minsky's financial instability hypothesis. The Post Keynesian approach to economic growth is shown to focus on the expansion of aggregate demand, with a distinctive approach to monetary, fiscal and other dimensions of macroeconomic policy. In conclusion the future prospects of Post Keynesian economics are assessed.

Keywords

Administered prices; American Economic Association; Animal spirits; Asset price inflation; Balance of payments constraint; Baran P. A.; Barter; Behavioural macroeconomics; Bounded rationality; Bubbles; Budget deficits; Business cycle theory; Cambridge capital controversies; Capital accumulation; Capitalism; Capital–labour substitution; Central banks; Class; Cost-push inflation; Critical realism; Deflation; Effective demand; Endogenous growth; Endogenous money; Entrepreneurship; Ergodicity and non-ergodicity in economics; European Central Bank; Evolutionary economics; Explanation; Financial regulation; Fiscal consolidation; Floating exchange rates; Functional finance; Fundamentalist Keynesians; General equilibrium; Government failure; Gross substitution; Harcourt G.; Harrod–Domar growth model; Hedging; Hicks J. R.; Hysteresis; Inequality; Income distribution; Income effects; Inflation; Institutional economics; International financial institutions; Investment decisions; Involuntary unemployment; IS–LM; Justice; Kahn R. F.; Kaldor N.; Kalecki M.; Keynes J. M.; Labour supply; Lending controls; Lexicographic preferences; Liquidity preference; Long run and short run; Lucas R.; Marginal productivity theory; Market failure; Market imperfections; Market power; Marx K. H.; Meade J. E.; Means G.; Methodology; Microfoundations; Minsky H.; Monetarism; Monetary policy; Money illusion; Myopia; National debt; Neoclassical growth theory; Neoclassical synthesis; Neutral money axiom; New Classical Economics; New Classical Macroeconomics; New Keynesian Economics; Open-system thinking; Paradox of costs; Paradox of thrift; Phillips curve; Political business cycles; Ponzi finance; Post Keynesian economics; Post-Walrasian theory; Price rigidity; Prices and incomes policies; Profit rate; Quantity theory of money; Rational expectations; Representative agents; Reserve requirements; Robinson J. V.; Savings propensities; Say's Law; Solow R.; Speculation; Sraffa P.; Stability and Growth Pact; Steindl J.; Stocks and flows; Structural adjustment; Structural change; Stylized facts; Substitution effects; Sweezy P. M.; Taylor rule; Trade unions; Transformational growth; Uncertainty; Wage rigidity; Washington Consensus; Weintraub, S.

JEL Classifications

B22; B59

Post Keynesian economics is a dissident school of macroeconomic thought based on a particular interpretation of John Maynard Keynes's *General Theory of Employment, Interest and Money* (1936).

Post Keynesian economics developed in the 1950s and 1960s in Cambridge (UK) and in the United States in the course of a critique of the so-called 'neoclassical synthesis' (sometimes also described as Old or Bastard Keynesianism). It represents both a recovery and an extension of Keynes's ideas (Palley 1996): a recovery, because Post Keynesians believe that the neoclassical interpretation of Keynes is profoundly misleading, and an extension, since they deal with important questions that Keynes neglected or ignored, including income distribution, social conflict, economic growth and inflation. Post Keynesian economics involves a distinctive approach to methodology, theory and policy (Holt and Pressman 2001; King 2003).

At the heart of Post Keynesian theory is the principle of effective demand, according to which output and employment are generally demand-constrained rather than supply-constrained. Post Keynesians claim to take the principle of effective demand more seriously do than mainstream macroeconomists, even those who describe themselves as 'Keynesians'. For Post Keynesians, demand constraints upon output and employment are not restricted to short period and are not the result of market imperfections or wage and price rigidities, but must be explained instead in terms of the characteristics of money and the pervasive influence of fundamental uncertainty. The six central messages of Keynes's vision may be summarized as follows. First, output and employment are determined in the product market, not in the labour market. Second, involuntary unemployment exists. Third, an increase in savings does not automatically generate an equivalent increase in investment. Fourth, a monetary economy is fundamentally different from a barter economy. Fifth, the quantity theory of money holds only under full employment, but cost-push forces may generate inflation well before this point is reached. Sixth, capitalist economies are driven by the 'animal spirits' of entrepreneurs, which determine the decision to invest (Thirlwall 1993).

It follows that Say's Law is false, and capitalism will normally not achieve or sustain full employment without government intervention. Post Keynesians therefore advocate the systematic use of fiscal and monetary policy to regulate aggregate demand, and deny the policy ineffectiveness propositions of mainstream macroeconomics. They advocate prices and incomes policies, rather than restrictive monetary policy, to control inflation.

A Brief Intellectual History

The origins of Post Keynesian economics may be traced back to the publication of the *General Theory* in 1936, since Keynes's masterpiece was open from the outset to alternative interpretations (King 2002). One of them, the IS–LM model developed by J. R. Hicks, James Meade and others, subsequently formed the core of the neoclassical synthesis model of output and employment in the short run. However, the Cambridge (UK) Post Keynesians, including Richard Kahn, Nicholas Kaldor, Joan Robinson and Piero Sraffa, directed their early criticisms against the long-run component of the neoclassical synthesis, the Solow growth model, in which full employment was ensured by capital–labour substitution along a well-behaved aggregate production function. The 'Cambridge capital controversies' of the late 1950s and early 1960s demonstrated the analytical failure of neoclassical growth theory, and were an important episode in the emergence of the Post Keynesian school (Mata 2004). Subsequently Robinson, Kaldor and the American Sidney Weintraub attacked the monetarist theory of inflation, emphasizing the causal role of the rate of change of money wages and arguing that monetary growth was the effect of inflation, not its cause. Kaldor, Weintraub and another American, Paul Davidson, were early advocates of the theory of endogenous money (Kaldor 1970).

Robinson conducted a lengthy correspondence with yet another American dissident, Alfred Eichner (Lee 2000). When, in December 1971, she gave the keynote Richard T. Ely lecture at the American Economic Association meeting in New Orleans to a large and enthusiastic audience, the defeat of the orthodox paradigm seemed to be only a matter of time (Robinson 1972). By the mid-1970s the term 'Post Keynesian' was widely used to describe the emerging school of thought (Eichner and Kregel 1975), which had broadened to include a systematic critique of the neoclassical synthesis. The IS–LM model was rejected, since uncertainty and animal spirits rendered the IS curve unstable, and endogenous money undermined the LM function. The Phillips curve model of wage inflation was rejected in favour of a socio-political analysis of distributional conflict and its resolution in a class society where capitalists enjoyed product market power and workers were highly unionized. And the marginal productivity theory of income distribution, discredited in the capital controversies, was replaced by a macroeconomic model that focused on the different savings propensities of capitalists and workers, or companies and households.

The critique of Chicago School monetarism was soon extended to New Classical Economics ('monetarism Mark II'). Post Keynesians objected to the principle of rational expectations, since it ignored the existence of fundamental uncertainty, and they denied the claim of Lucas and his associates that fiscal and monetary policy could have no effect on output or employment. They saw only slightly more merit in New Keynesian

Economics, with its emphasis on market imperfections as the source of all macroeconomic problems. More was involved than disagreements on economic theory; important methodological and policy issues were also at the heart of these criticisms.

The mainstream never accepted the Post Keynesian critique. Orthodox Keynesian macroeconomists like Robert Solow accused the Post Keynesians of incoherence; they were united, he suggested, only by what they were against. Insiders distinguished three Post Keynesian schools, the Kaleckians, the Sraffians and the 'fundamentalist Keynesians', with a number of prominent individualists who belonged to none of them (Harcourt 1987; Hamouda and Harcourt 1988). Divisions remain on the respective virtues of a 'big tent' and a 'small tent' definition of Post Keynesianism.

In 2006 Post Keynesians were a small, embattled minority, strongest in France, Italy and a few institutions in the United States (especially the University of Missouri at Kansas City), with outposts in Britain and Australia. They published in a range of heterodox journals, especially in the *Journal of Post Keynesian Economics*, founded by Davidson and Weintraub in 1977, in the *Cambridge Journal of Economics, Journal of Economic Issues and Review of Political Economy.*

Methodology

As is common with dissenting schools of thought, Post Keynesians have always recognized the importance of methodology. They have been strongly influenced by Keynes's philosophical writings (O'Donnell 1989), identifying with his insistence on 'open-system thinking' and organic rather than atomistic models of human behaviour, his distrust of formalism (and of econometric modelling in particular), and his belief that economics was nothing if not a policy science – or an art, perhaps.

On many specific methodological questions, Post Keynesians stress their differences with the mainstream (Dow 1996). They doubt the relevance of equilibrium models, which fail to allow for cumulative causation, the role of history or the importance of hysteresis. Their emphasis on uncertainty leads them to reject the 'rational expectations' principle and to assert the importance of habit, convention and social institutions in the formation of business expectations. Post Keynesians criticize the mainstream insistence that 'microfoundations' must always be provided for macroeconomic theory, because this denies the existence of emergent properties of macro systems that cannot be inferred from their micro components, and thus involves a fallacy of composition. Microeconomic theory needs *macrofoundations*, they maintain. Finally, Post Keynesians take a quite distinctive approach to long-run theory. Since the principle of effective demand applies in the long run, no less than in the short run, Post Keynesian growth theory stresses the role of demand as a determinant of economic growth, and does not impose a condition that resources (including labour) are always fully employed, and output constrained solely by supply, in the long run.

Many (though not all) Post Keynesians are attracted by *critical realism* as a unifying methodological position (Lawson 2003). There are many points of contact, including the critical realists' endorsement of open-system thinking; their denial of 'event regularities' of the type needed if standard econometric estimation techniques are to be generally reliable; and their stress on the importance of ontology and the identification of causal processes and mechanisms as the key to explanation in social science. Critical realism has become a significant point of contact between Post Keynesians and other schools of heterodox economic thought.

Macroeconomic Theory: The Short Period

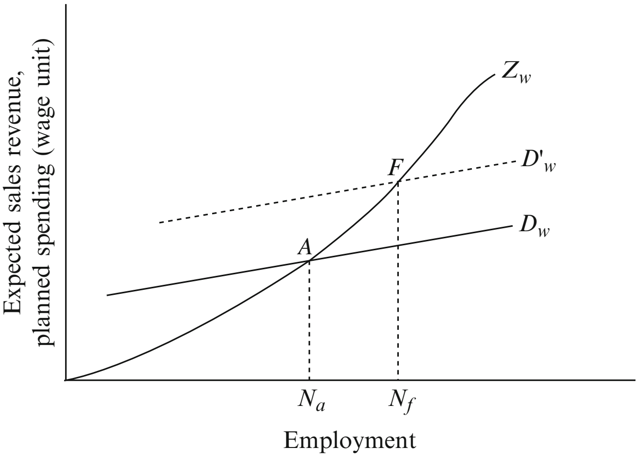
The short-period theory of output and employment is the core of Post Keynesian macroeconomics. In the short period, the capital stock is held constant. This is done purely for analytical convenience; there is no presumption that the theory of effective demand is irrelevant to the long period, when the accumulation of capital is brought into the analysis. Thus the Post Keynesian treatment of the short and long periods must be distinguished from the neoclassical analysis of the 'short run' (in which demand matters) and the 'long run' (when it does not).

There is no single canonical short-period Post Keynesian model. The three most influential models are those of Paul Davidson, Michal Kalecki and Hyman Minsky, which differ in some important respects. They are not, however, entirely incompatible. All agree on the central role of the principle of effective demand; the defects of the IS-LM model; the importance of uncertainty, money and finance (this is largely implicit in the Kalecki version, and quite explicit in the other two); the consequent repudiation of rational expectations; and the policy implications, which include the need for government intervention to stabilize the economy and to maintain full employment. They differ on some questions of microeconomics (there is no question of providing microfoundations), on the detailed treatment of money and finance, and most obviously on the social and political context, above all on the class-driven or class-blind nature of the analysis.

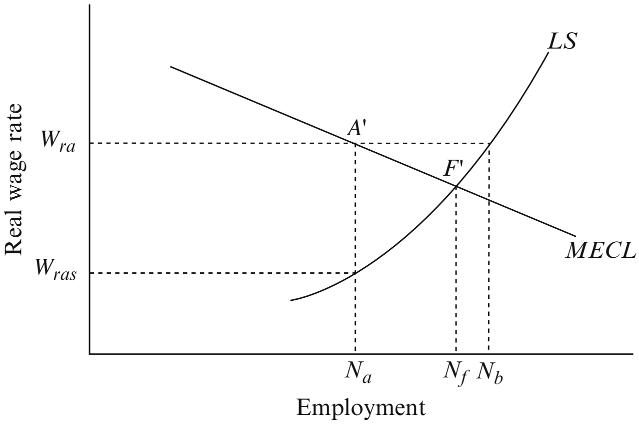
The Fundamentalist Keynesian Model

This is an elaboration of the aggregate supply-aggregate demand model set out by Keynes himself in the early chapters of the *General Theory* (Keynes 1936, ch. 3). (Note that this is emphatically *not* the textbook model in price level/real output space, which is a teaching version of the neoclassical synthesis and would be repudiated by all Post Keynesians.) Originating in the 1950s with Weintraub, it has been propagated tirelessly over several decades by Davidson in a series of books beginning with *Money and the Real World* in 1973 and culminating in *Financial Markets, Money and the Real World* (Davidson 2002). In Fig. 1 the aggregate supply function Z_w links total employment to total expected sales; it slopes upwards since total costs of production (including gross profits) increase as employment rises. The aggregate demand function D_w does not coincide with Z_w, as it would in an economy where Say's Law prevails. It, too, slopes upwards, as planned spending also rises with employment. The point of effective demand is A, where the two curves intersect, and aggregate employment is given by N_a. The labour market implications are illustrated in Fig. 2 (which, like Fig. 1, comes from Davidson 1999, not directly from Keynes). Employment is determined in the product market, in Fig. 1. The real wage can be established from the market equilibrium curve of labour-hire (MECL) in Fig. 2; it is W_{ra}. Davidson emphasizes that MECL is *not* the labour demand curve; employment depends on aggregate demand and is therefore determined in the product market, not the labour market. Involuntary unemployment is N_aN_b; it is not due to (real or money) wage rigidity, but to deficient effective demand.

Post Keynesian Economics, Fig. 1 Aggregate supply and demand (*Source:* Davidson (1999, p. 582))



Post Keynesian Economics, Fig. 2 Wages and employment (*Source:* Davidson (1999, p. 582))



For full employment to be achieved, the aggregate demand curve would need to shift to D'_w, increasing employment to N_f, with a movement from A to F in Fig. 1 and a corresponding move from A' to F in Fig. 2. Without such a shift there will be no increase in employment, no matter how willing workers might be to accept a cut in either money or real wages. In fact Fig. 2 reveals that W_{ras} is the reservation wage of the marginal unemployed worker, but in the absence of an adequate level of effective demand this is simply not relevant.

Underlying Keynes's model, Davidson argues, is a rejection of the three fundamental axioms of mainstream macroeconomics. The *axiom of ergodicity* asserts that the future can be reliably inferred from the past. The *axiom of gross substitution* asserts that flexibility in relative prices will ensure that all markets clear. The *neutral money axiom* ensures that changes in the stock of money have no permanent effects on real output or employment. Non-ergodicity creates radical uncertainty, which induces people to hold money; since goods are not perfect substitutes for money, money is not neutral, even in the long run, and Say's Law is false. The non-neutrality of money does not require 'money illusion' on the part of any agent. Involuntary unemployment is not the result of wage or price rigidity, and can be eliminated only by increases in effective demand. Wage reductions will prove futile, or even counter-productive (since deflation depresses business and household confidence and increases real interest rates).

The Kaleckian Model

Michal Kalecki discovered the principle of effective demand in the 1930s independently of Keynes under the influence of Rosa Luxemburg and, through her, of Karl Marx. The class distinction between capitalists and workers, which is only implicit in the *General Theory,* occupies centre stage in Kalecki's analysis. In place of a single consumption function, there are two. Workers save nothing from their wages, while capitalists save a constant proportion of their profit income. Kalecki's famous aphorism that 'capitalists get what they spend, while workers spend what they get', can be derived from the simple income-expenditure model set out in his 1939 *Essays in the Theory of Economic Fluctuations* (1990, pp. 233–318), itself an elaboration of his 1933 model of the business cycle (1990, pp. 65–108). In the simplest case, if we neglect both the government and the foreign sector, total wage (W) and profit (P) income is equal to the sum of consumption expenditure by workers (C_w) and capitalists (C_c) plus investment spending (I):

$$W + P = C_w + C_c + I. \tag{1}$$

Since $W = C_w$ by assumption, it follows that

$$P = C_c + I, \tag{2}$$

so that profits are equal to the sum of capitalists' consumption and investment expenditure. If investment is a positive function of expected profits, which are themselves closely related to recent past profits, this leads directly to a demand-driven model of the trade cycle. If we incorporate the government and overseas sectors, eq. (1) can be replaced with

$$W + P + T + M = C_w + C_c + I + G + X \quad (3)$$

where taxes (T) and imports (M) are added to the income side of the equation and government expenditure (G) and exports (X) to the expenditure side. It follows that

$$\mathrm{P} = C_c + I + (G - T) + (X - M), \quad (4)$$

so that capitalists profit from both government deficits ($G – T$) and trade surpluses ($X – M$).

Unlike Keynes, Kalecki had no time for the marginal productivity theory of distribution. In his model the share of profits in total output is determined by the degree of monopoly in the product market. Outside agriculture, oligopoly rather than perfect competition is the rule. Firms set prices by marking up their average variable costs of production, the markup varying inversely with the degree of competition that they face. This, Kalecki argues, establishes a strong tendency for a chronic deficiency in effective demand, since the wage share will normally be too small (and the profit share too high) to generate enough consumption expenditure to maintain full employment. This aspect of Kalecki's analysis was emphasized by 'left Keynesians' like Josef Steindl and the neo-Marxists Paul Baran and Paul Sweezy in their work on *monopoly capital.*

Kalecki also highlights the class nature of capitalist society in the context of macroeconomic policy. Capitalists will resist deficit-financed spending by the government, even though it might be expected to increase total profits. This is only partly due to an unthinking attachment (encouraged by orthodox economists) to the principles of sound finance. It has more rational roots in their concern to avoid competition from state-owned enterprises, and more especially in their well-founded fear that full employment will prove inconsistent with 'discipline in the factories'. As early as 1943 Kalecki was predicting the emergence of a *political business cycle,* in which fiscal and monetary policy is repeatedly eased before elections and tightened (under pressure from business interests) soon afterwards (Kalecki 1990, pp. 347–56). In the 1950s, viewing the Cold War from his native Poland, he criticized the 'military Keynesianism' of Western governments – capitalists had welcomed demand-boosting armaments spending while resisting more socially useful civilian expenditures.

The Financial Instability Hypothesis

The Kaleckian model is characterized by a relative neglect of money and finance. Hyman Minsky's version of the short-period Post Keynesian model quite explicitly aimed to put finance back into business cycle theory (Minsky 1986). His 'Wall Street vision' of capitalism focuses on the relationship between investment bankers and their customers, by contrast with the 'village fair' conception of exchange between individual small producers that underpins mainstream theory. Like Davidson and Kalecki, Minsky sees fluctuations in investment expenditure as the principal cause of economic fluctuations. The investment decisions of capitalists are constrained by their ability to pay for them, and this is conditioned by lenders' estimates of their ability to repay. Minsky distinguishes three phases of the cycle. In the immediate aftermath of a crisis lenders are very cautious, and accommodate only those borrowers who can demonstrate an ability to service their loans and repay the principal on time; this is the phase of *hedge finance.* As the upswing gathers pace, and memories of previous difficulties begin to fade, it becomes possible to borrow for more questionable projects, where interest payments are covered by expected profits, but not repayments of principal; this is the phase of *speculative finance.* In the final stages of the boom caution is thrown to the winds and lenders now provide *Ponzi finance* (the term is derived from a notorious early 20th-century swindler): new borrowing is now required to enable borrowers to make interest payments on previous loans. When lenders' confidence collapses, borrowers are unable to obtain refinance and are forced to sell securities and other assets at 'fire sale' prices. In the ensuing financial crisis, real investment falls and the economy moves into recession. The early stages of recovery are again characterized by the provision of hedge finance, and so the cycle repeats itself, over and over again. Memories are short, and expectations are far from rational.

Rather late in his career Minsky discovered Kalecki, and added the Kaleckian theory of profits to his own model. This gave him a theory of firms' financial resources to set against his original analysis of their financial commitments, and reinforced the policy implications that he had drawn from the financial instability hypothesis. It can be seen from eq. (4) that aggregate profits are increased by higher budget deficits, and reduced by fiscal conservatism. Deficits, then, are good for business. There is a stock dimension as well as a flow dimension to this conclusion. The United States financial system was much less fragile after 1945 than it had been in 1929, Minsky argued, in large part because of the cumulative impact of wartime and post-war deficits. The huge growth in the federal government debt had provided the private sector with massive quantities of risk-free government securities, thereby rendering their asset portfolios much more robust than they had previously been. Minsky was therefore a supporter of big government. He also advocated tight and intrusive regulation of financial markets, and argued that central banks should recognize their duty to act as lender of last resort to Wall Street, no matter how much this increased the dangers of moral hazard. But he doubted whether the inherent instability of the capitalist economy could ever be completely overcome.

Some Comparisons

The similarities between these three models are much more important than their differences, especially when they are contrasted with the 'new consensus' model of mainstream macroeconomics. There are no 'microfoundations', and certainly no attempt is made to ground the analysis in any form of multi-period utility-maximizing model of general equilibrium under rational expectations. This is ruled out by the non-ergodicity axiom in the Fundamentalist Keynesian model, and by the cyclical myopia of borrowers and lending institutions that is central to the financial instability hypothesis (Kalecki's analysis of 'lender's risk' and 'borrower's risk' has affinities both with Minsky and with Keynes's treatment of fundamental uncertainty). There are no 'representative agents': capitalists and workers in Kalecki, borrowers and lenders in Minsky (and bulls and bears in Keynes's analysis of liquidity preference) are structurally and behaviourally heterogeneous. Deflation is viewed as part of the problem – a very important part, at least for Minsky – not as the solution to macroeconomic difficulties. Cyclical fluctuations originate in the private sector, due to the volatility of business investment decisions, not in the policy errors of the public sector. And government intervention is essential to 'stabilize an unstable economy', to paraphrase the title of Minsky's last (1986) book – though neither he nor Kalecki minimized the obstacles that it would encounter.

These are not the only short-period Post Keynesian models, though they remain the most influential. They are, it must be repeated, all inconsistent with the 'new consensus' in macroeconomics, which can be encapsulated in three equations. Post Keynesians dispute all three. They reject the aggregate demand curve, on the grounds that interest rates are less important, and uncertainty-induced shifts in the curve much more important, than the mainstream is willing to admit. They are equally critical of the Phillips curve, since it neglects socio-political institutions and denies any role for class conflict over income distribution. And they criticize the Taylor rule that underpins the monetary policy response function, as it uses a single instrument (the short-term interest rate) instead of many to influence the wrong objective (output price inflation instead of employment, neglecting asset price inflation). More will be said about Post Keynesian thinking on money and inflation in a later section.

Macroeconomic Theory: The Long Period

In the *General Theory* Keynes analysed the effects of investment in a short-period model in which, by definition and purely as a simplifying assumption, the capital stock was held constant. 'Generalising the *General Theory*' to the long period, Post Keynesian theories of capital accumulation take as their starting point the Harrod–Domar growth model (which Kalecki extended further to apply to socialist economies). There is no requirement that capital or labour are fully employed or that the growth path will be stable; this can be expected, as Robinson put it in her *Accumulation of Capital* (1956), only in a mythical 'golden age'. The Cambridge capital controversies demonstrated that the neoclassical adjustment mechanism – capital–labour substitution in response to changes in relative factor prices – is not in general a reliable one. Differences between the actual, equilibrium (or 'warranted') and maximum possible (or 'natural') rates of growth might be eliminated through changes in the average propensity to save induced by changes in income distribution. But, Post Keynesians maintain, there are no grounds for supposing that effective demand is unimportant in the long period, or for the neoclassical belief that economic growth is entirely supply-determined.

There are a number of Post Keynesian models of demand-driven growth (Setterfield 2002). All of them invoke 'Say's Law in reverse', according to which aggregate supply (and potential output) responds to the growth of aggregate demand (and actual output). *Kaldorian* models treat exports as the only truly exogenous source of demand and highlight the balance of payments constraint on economic growth, which is especially (but not exclusively) relevant to developing economies. *Kaleckian* models focus on the connection between wages, consumption and aggregate demand, adding to the familiar *paradox of thrift* (in which an increased propensity to save reduces income and keeps the volume of saving unchanged) a *paradox of costs,* in which an increase in the real wage increases workers' consumption, raises the level of capacity utilization and thereby leads to a higher rate of profit. Finally, there are models of *transformational growth* associated with Luigi Pasinetti and Edward Nell, in which capital accumulation is inextricably linked to structural change. Once again attention is concentrated on demand conditions; this time, however, it is investment demand that plays the crucial role.

These Post Keynesian growth models are all radically different from neoclassical theories, including both the canonical Solow model and the more recent 'New' or 'endogenous growth' models (though they share with the latter a denial of diminishing returns in the manufacturing and advanced service sectors). The Post Keynesians assert the continuing importance of the principle of effective demand and the irrelevance or reversal of Say's Law, since in the long period demand tends to create its own supply. They have no truck with marginal productivity theory or with the use of aggregate production functions of any description.

There are connections between Post Keynesian growth theory and the treatment of capital accumulation in other heterodox traditions, especially the radical-Marxian focus upon the class nature of capitalist society, the critical role of the profit rate and the instability of the capitalist growth path. Equally, the emphasis placed in evolutionary and Schumpeterian theory on the role of entrepreneurs, the importance of finance and the cyclical nature of growth is fully consistent with the Post Keynesian approach.

Post Keynesian Microeconomics

Post Keynesian microeconomics is relatively underdeveloped. There are methodological reasons for this, since (as we have seen) Post Keynesians reject the neoclassical requirement that rigorous microfoundations be provided for macroeconomic theory. Although microeconomics is not needed as the basis for serious macroeconomic thinking, Post Keynesians are nevertheless highly critical of many aspects of mainstream microeconomic analysis, including the modelling of equilibrium, the elimination of uncertainty by expressing all relevant magnitudes in certainty-equivalents, and the reliance on identical or 'representative' rather than heterogeneous agents.

As in macroeconomics, in their microeconomics Post Keynesians are concerned with the real world, and insist that formal models must bear a close relation to the 'stylized facts' of modern capitalism. Thus Post Keynesian pricing theory (Lee 1998) addresses itself to the large oligopolistic corporation, not to an imaginary world of perfect competitors. Drawing on the work of Kalecki, Philip Andrews, Gardiner Means, Paolo Sylos-Labini and Alfred Eichner, it models the formation of *administered prices,* with firms first adding a markup to their variable costs of production and then selling as much as they can given the prevailing demand conditions. Prices increase only if costs rise, or under quite exceptional demand pressure. This also provides a Post Keynesian theory of income distribution, since the average degree of monopoly, which determines markups, is the most important determinant of the income shares of wages and profits. Changes in the degree of monopoly have other macroeconomic consequences, for both inflation and aggregate demand.

Post Keynesian consumer theory is still in its infancy. It places more emphasis on income effects than on substitution effects and replaces the neoclassical axioms of rational choice with a theory of lexicographic preferences where habit, custom and social convention are important constraints on individual behaviour (Lavoie 1992, pp. 61–92). The full implications for labour supply decisions have yet to be fully worked out. In their microeconomics Post Keynesians have drawn heavily on insights from other more or less heterodox schools of thought, especially institutional and evolutionary economics. Much remains to be done to extend and deepen Keynes's early exploration of the role of habit, conventions, bounded rationality and rules of thumb in individual and corporate decision-making. The lack of a distinctive Post Keynesian welfare economics is a particularly important weakness, which has hindered the emergence of a coherent approach to environmental issues (Winnett 2003).

Economic Policy

There is, however, a very clear Post Keynesian position on matters of macroeconomic policy. Since Say's Law is rejected, in both the short period and the long period, the principle of effective demand is the foundation for monetary and fiscal policy. This leads Post Keynesians to a broadly social democratic position, not far removed from that of the Old Keynesian advocates of the 1950s–1960s neoclassical synthesis. Thus they favour big government, since it is more likely than small government to be able to stabilize the level of economic activity and achieve full employment. Post Keynesians worry much less about state failure than about market failure. Unlike both Old and New Keynesians, however, they are very clear that market imperfections, and the associated wage and price rigidities, are not at the root of macroeconomic problems. There is no point in using an imaginary world of perfect competition as a reference point. Deflation, even if it were practicable, would be undesirable and counter-productive. Increased inequality is also likely to have adverse macroeconomic consequences, notably if the Kaleckian 'paradox of costs' applies, so that stabilization policy need not conflict with the imperatives of social justice. Post Keynesian rejection of neoliberal policies carries over to a comprehensive critique of the 'Washington Consensus' on policy for developing countries. At the same time Post Keynesians are not Stalinists; they aim to make markets work better, not to eliminate them. This, they argue, requires wide-ranging government intervention, with a number of macroeconomic targets and a variety of instruments.

Monetary Policy

Post Keynesian thinking on monetary policy developed out of opposition to monetarism in the early 1970s and to the policy prescriptions of New Classical macroeconomics in later years. Post Keynesians insisted that, since money was endogenous, the stock of money was not a control variable or a feasible policy instrument. Thus monetary policy must necessarily operate via central bank control over the (short-term) rate of interest, and would inevitably have consequences for output and employment as well as for the inflation rate. In this they were proved to be entirely correct, and they are entitled to view the treatment of monetary policy in the 'new consensus' as a vindication of the Post Keynesian critique. However, they also criticize the Taylor rule on the grounds that it is aimed at the wrong target and relies upon a single, very blunt instrument. There is a strong case, they argue, for reviving prices and incomes policy to combat the danger of inflation, using monetary policy to target and output and employment, asset price inflation and financial fragility. Post Keynesians regard stock market and housing bubbles, and rising levels of household and corporate debt, as serious problems that need policy solutions. No single-instrument approach to these problems will succeed. Alternative instruments include the (re)introduction of direct controls over lending, the tightening of financial regulations, and more market-friendly measures such as asset-based reserve requirements that would operate as a tax on types of lending that the authorities wished to discourage. Post Keynesians have been particularly critical of the highinterest policy adopted by the European Central Bank and the apparent absence of a lender of last resort within the Eurozone.

Fiscal Policy

Post Keynesians are no less critical of recent mainstream thinking on fiscal policy. Here the Old Keynesian principle of *functional finance* has given way to the pre-Keynesian principle of *sound finance,* which is invariably interpreted as requiring balanced budgets in the short run and fiscal consolidation (budget surpluses, and a reduction in government debt) in the long run. For Post Keynesians, the principle of effective demand should govern fiscal policy, and governments should run deficits, or surpluses, or (exceptionally) balanced budgets, depending solely on the macroeconomic requirement of achieving full employment with an acceptably low inflation rate. Some would go further in the direction of 'unsound finance', since in the Kaleckian short-period model permanent deficits mean permanently higher business profits and hence higher – not lower – levels of investment expenditure; private spending is *crowded in* by government spending, not crowded out. Hyman Minsky added a stock dimension to this argument: the government debt accumulated as the sum of past deficits serves to render private sector balance sheets more robust and thus to reduce the danger of financial instability. The European Union's 'Stability and Growth pact' is in fact a recipe, Post Keynesians argue, for stagnation and instability.

Prices and Incomes Policy

Post Keynesians deplore the way in which one entire area of macroeconomic policy has disappeared completely from the mainstream agenda: prices and incomes policy is no longer taken seriously as an anti-inflationary instrument, and indeed is rarely discussed. In the 1960s and 1970s Post Keynesians led the way in proposing innovative alternatives to deflationary monetary and fiscal policies, including the tax-based incomes policy proposed by Wallich and Weintraub (1971) and more centralized neocorporatist social agreements that worked well for many years in much of northern Europe (Cornwall and Cornwall 2003). Weaker unions, reduced levels of industrial conflict and much lower rates of wage and price inflation have reduced the attraction of measures such as these, but Post Keynesians regard them as a valuable policy resource should the inflation dragon once more rear its ugly head.

Other Policy Issues

On questions of international economic policy Post Keynesians tend to be sceptical of the benefits of unregulated free trade and free capital movements (Blecker 2003). Many regard floating exchange rates as a major source of macroeconomic instability, urging instead a reconsideration of Keynes's ambitious plans for an International Clearing Union (Davidson 2002; Milberg 2003; Vernengo 2003). For the long period they focus on demand-side policies for economic growth and criticize the deflationary bias of the structural adjustment programmes imposed on developing countries by the international financial institutions. As already noted, there is no specifically Post Keynesian welfare economics, so that there is also no genuinely distinctive position on most microeconomic issues (including environmental questions, industry policy, labour market regulation and antitrust). However, their social democratic sympathies do tend to bring Post Keynesians close to the institutionalist or 'left neoclassical' position on many of these questions.

Assessment and Prospects

In the 1970s many Post Keynesians believed that mainstream economics was in a state of Kuhnian crisis, with the very real prospect of a paradigm shift in their favour. This confidence proved to be misplaced, and by the first decade of the 21st century Post Keynesian economics had been thoroughly marginalized. This had something to be with the sociology of the profession, which displayed an increasing intolerance of alternative perspectives and methods of research. Some Post Keynesians suspected that they had to share part of the blame, since there was some truth in the accusation that the Post Keynesian church had become too broad, with a message that was lacking in coherence. Part of the problem was that the mainstream itself had changed, with the New Keynesians adopting some Post Keynesian positions (for example on endogenous money and the consequent rejection of the LM schedule) while rejecting many others. New Keynesian economics offered a rather less clear target for Post Keynesian criticism than the neoclassical synthesis, or monetarism, or New Classical economics. Subsequent developments in 'behavioural macroeconomics' and 'post-Walrasian theory' confused the picture still further, making the oppositional character of Post Keynesian economics less easy to define. The future relationship between Post Keynesianism and heterodox economics more generally was also unclear, with familiar fault lines developing here, too, between sectarians and pluralists. For all that, Thirlwall's (1993) six propositions with which this entry began are clear enough, and important enough, to provide a secure intellectual future for this unusually persistent and perceptive dissenting minority.

See Also

- ▶ Functional Finance
- ▶ Heterodox Economics
- ▶ Keynesianism
- ▶ Macroeconomics, Origins and History of
- ▶ Money
- ▶ Unemployment

Bibliography

Blecker, R.A. 2003. International economics. In *The Elgar companion to post Keynesian economics*, ed. J. King. Cheltenham: Elgar.

Cornwall, J., and W. Cornwall. 2003. *Capitalist development in the twentieth century: An evolutionary Keynesian analysis*. Cambridge: Cambridge University Press.

Davidson, P. 1999. Keynes' principle of effective demand versus the bedlam of the New Keynesians. *Journal of Post Keynesian Economics* 21: 571–588.

Davidson, P. 2002. *Financial markets, money, and the real world*. Cheltenham: Elgar.

Dow, S. 1996. *The methodology of macroeconomic thought*. Cheltenham: Elgar.

Eichner, A., and J. Kregel. 1975. An essay on post Keynesian theory: A new paradigm in economics. *Journal of Economic Literature* 13: 1293–1314.

Hamouda, O., and G. Harcourt. 1988. Post Keynesianism: From criticism to coherence. *Bulletin of Economic Research* 40: 1–33.

Harcourt, G. 1987. Post–Keynesian economics. In *The new palgrave: A dictionary of economics*, ed. J. Eatwell, M. Milgate, and P. Newman, vol. 3. London: Macmillan.

Holt, R., and S. Pressman, eds. 2001. *A new guide to post Keynesian economics*. London: Routledge.

Kaldor, N. 1970. The new monetarism. *Lloyd's Bank Review* 97 (July): 1–18.

Kalecki, M. 1990. *Collected works of Michal Kalecki. Volume I: Capitalism: Business cycles and full employment*. Oxford: Clarendon Press.

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

King, J. 2002. *A history of post Keynesian economics since 1936*. Cheltenham: Elgar.

King, J., ed. 2003. *The Elgar companion to post Keynesian economics*. Cheltenham: Elgar.

Lavoie, M. 1992. *Foundations of post Keynesian economic analysis*. Aldershot: Elgar.

Lawson, T. 2003. *Reorienting economics*. London: Routledge.

Lee, F. 1998. *Post Keynesian price theory*. Cambridge: Cambridge University Press.

Lee, F. 2000. On the genesis of post Keynesian economics: Alfred S. Eichner, Joan Robinson and the founding of post Keynesian economics. In *Research in the history of economic thought and methodology*, ed. W.J. Samuels, vol. 18c, 1–258. Amsterdam: JAI/Elsevier.

Mata, T. 2004. Constructing identity: The post Keynesians and the capital controversies. *Journal of the History of Economic Thought* 20: 241–260.

Milberg, W. 2003. Globalization. In *The Elgar companion to post Keynesian economics*, ed. J. King. Cheltenham: Elgar.

Minsky, H. 1986. *Stabilizing an unstable economy*. New Haven: Yale University Press.

O'Donnell, R. 1989. *Keynes: Philosophy, politics and economics*. London: Macmillan.

Palley, T. 1996. *Post Keynesian economics*. London: Macmillan.

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

Robinson, J. 1972. The second crisis of economic theory. *American Economic Review* 62 (2): 1–10.

Setterfield, M., ed. 2002. *The economics of demand-led growth*. Cheltenham: Elgar.

Thirlwall, A. 1993. The renaissance of Keynesian economics. *Banca Nazionale del Lavoro Quarterly Review* 186: 327–337.

Vernengo, M. 2003. Bretton Woods. In *The Elgar companion to post Keynesian economics*, ed. J. King. Cheltenham: Elgar.

Wallich, H., and S. Weintraub. 1971. A tax-based incomes policy. *Journal of Economic Issues* 5: 1–19.

Winnett, A. 2003. Environmental economics. In *The Elgar companion to post Keynesian economics*, ed. J. King. Cheltenham: Elgar.

Postan, Michael Moïssey (1899–1981)

Robert Brenner

Keywords

Economic demography; Feudalism; Peasants; Postan, M. M.; Serfdom

JEL Classifications

B31

M.M. Postan, who was born in Tighina, Bessarabia, in 1899 and who died in Cambridge in 1981, was one of most distinguished economic historians of the 20th century. After briefly studying natural sciences and sociology at the University of St Petersburg he moved on to study law and economics at the University of Odessa, and then at the University of Kiev. He came to England in 1920 and between 1921 and 1926 took his first degree, his MA, and his Ph.D. at the London School of Economics. Between 1927 and 1937 he held lectureships, successively, at University College, London, at the London School of Economics, and at Cambridge University. In 1938 he was appointed to succeed Sir John Clapham in the chair of economic history at Cambridge, a position he retained until his retirement.

A specialist in medieval economic history, Postan originally made his reputation during the late 1920s and early 1930s on the basis of his studies on medieval trade and finance. His joint volume, with Eileen Power, Studies in English Trade in the Fifteenth Century (1933), became a standard work. He also published such seminal articles as ‘Credit in Medieval Trade’ (1928) and ‘Recent Trends in the Accumulation of Capital’ (1935).

From the later 1930s, Postan began to present his own distinctive interpretation of long-term trends in the medieval economy. In ‘The Chronology of Labour Services’ (1937) and ‘The Rise of the Money Economy' (1944) Postan advanced devastating critiques of the hitherto-dominant unilineal evolutionist interpretation of pre-industrial European economic development, as advanced by distinguished medievalists such as Henri Pirenne. According to that view, it was the more or less steady expansion of commerce which drove the European economy forward, leading first to the decline of serfdom, next to the differentiation of the peasantry and the rise of agrarian capitalism, and ultimately to the growth of manufacturing and modern industry. Postan showed, in contrast, that trade, in itself, could as easily lead to the strengthening of the old, pre-capitalist forms as to their dissolution. He illustrated his point with a detailed discussion of the fluctuations of labour services in medieval England, showing that they rose and fell in direct proportion to lordly demands for labour and, in particular, that demesne production grew, serfdom was intensified and labour services increased in precisely those areas of the country which were most commercialized and most exposed to the London market. Postan also pointed out that the spectacular rise of serfdom in late medieval and early modern Europe took place in large part in response to the rise of the international grain market.

In his 1950 report to the Ninth Congress of Historical Sciences, Postan put forward the initial version of his own population-centred interpretation of medieval economic history as an alternative to the trade-centred interpretation. In this and later work, Postan demonstrated that the pre-industrial economy of Europe was marked by a succession of long cycles of demographically driven expansions and contractions, following a basically Malthusian dynamic. He then went on to argue, in Ricardian fashion, that during the up phase of these cycles declining returns in agriculture (declining productivity) determined rising rents, falling wages, and terms of trade running in favour of agricultural and against industrial goods, while in the down phase, rising returns in agriculture determined just the opposite trends. Postan's interpretation followed lines which had begun to be sketched by the German demographic historian Wilhelm Abel and it influenced, in turn, the work of the French agrarian historian of the early modern period, Emmanuel Le Roy Ladurie. By the later 1950s, Postan's demographic view already had been so widely accepted as the key to the interpretation of pre-industrial economic change, that H.J. Habakkuk could reasonably conclude, in a synthetic essay on 'The Economic History of Modern Britain' for the *Journal of Economic History* in 1958, that

> For those who care for the overmastering pattern, the elements are evidently there for a heroically simplified version of English history before the nineteenth century in which the long-term movements in prices, in income distribution, in real wages, and in migration are dominated by changes in the growth of population.

Postan further developed his interpretation in a long series of specialized studies on all aspects of the medieval economy – agricultural technique, agricultural investment, the legal status of the peasantry, and so on – as well as in a number of major syntheses. In all these works, he remained guided by the conviction that the best results would come by linking, as closely as possible, generalizations derived from economic theory with the results of exhaustive primary research.

See Also

► Feudalism

Selected Works

1928. Credit in medieval trade. *Economic History Review* 1: 234–261.

1933. (With E. Power.) *Studies in English trade in the fifteenth century.* London: G. Routledge & Sons.

1935. Recent trends in the accumulation of capital. *Economic History Review* 6(1): 1–12.

1937. The chronology of labour services. *Transactions of the Royal Historical Society, 4th series* 20: 169–193.

1939. *The historical method in social science*. London: London University Press.

1944. The rise of the money economy. *Economic History Review* 14.

1954. *The famulus: The estate labourer in the XIIth and XIIIth centuries*. London/New York: Cambridge University Press.

1966. Medieval agrarian society in its prime: England. In *The Cambridge economic history of Europe,* vol. 1, *The agrarian life of the Middle Ages,* 2nd ed., ed. M.M. Postan and H.J. Habakkuk. Cambridge: Cambridge University Press.

1971. *Facts and relevance: Essays on historical method*. London: Cambridge University Press.

1972. *The medieval economy and society: An economic history of Britain in the Middle Ages*. London: Weidenfeld & Nicolson.

1973a. *Medieval trade and finance: Collected essay.* Cambridge: Cambridge University Press.

1973b. *Essays on medieval agriculture and general problems of the medieval economy.* London: Cambridge University Press.

Bibliography

Habakkuk, H.J. 1958. The economic history of modern Britain. *Journal of Economic History* 18: 486–501.

Postlethwayt, Malachy (1707–1767)

William Darity Jr.

Malachy Postlethwayt gave vent to the most comprehensive expression of mercantilist thought on behalf of British imperial interests. Fay (1934, p. 3) justifiably called Postlethwayt, alongside Joshua Gee, a major 'spokesman' for 18th-century England. Postlethwayt's mercantilist vision emphasized (1) the slave trade to Africa and slavery in the Caribbean as vital stimuli to development of British manufactures; (2) the Royal African Company as an instrument of management of 'the African trade'; (3) the necessity of competition with France for control of the slave trade; and (4) the general principle that government must promote trade and industry.

His monumental *Universal Dictionary of Trade and Commerce*, 20 years in the making before its first edition was published in instalments over the interval 1751–55, included an entry entitled 'Africa', summarizing his views on the relationship between African slavery and British industry. Despite acknowledging the brutality of the trade and allusion to some future date when a 'Christian spirit' might be moved to end the trade, Postlethwayt was wholly pragmatic. After all, he concluded, the gains for Britain from the slave trade were substantial – being a 'trade (that) is . . . all profit' and a trade that 'occasionally gives so prodigious employment to our people both by sea and land'.

This perspective resonated throughout Postlethwayt's pamphlets (see his *Selected Works*). Sir James Steuart may have been the 'last' British mercantilist, but he certainly was not the purest. For that we must turn to Postlethwayt, whose vision was undiluted by vestiges of humanitarism.

Although foreign trade, with the slave trade as a key component, was Britain's engine of growth for Postlethwayt, there was great breadth in the matters he viewed as relevant to British economic development. Scientific and technical advances, maintenance of low or zero interest rates (see Viner, 1937, p. 47), sport and leisure (Dorfman, 1971, p. 7), the public debt (Johnson, 1937, pp. 190–5), agricultural policy (Johnson, 1937, pp. 196–201), maintenance of low wages, and development of securities markets were among the many factors he identified as influences on the rate of economic expansion. Nevertheless, the overseas 'plantations' or 'colonies' lay at the heart of Postlethwayt's mercantile system, and, for Postlethwayt, full development of the plantations required slaves. Indeed, Postlethwayt's writings provided compelling evidence for Eric Williams's view in *Capitalism and Slavery* (1944) that British mercantile strategists were aware of slave-trading and slavery's ramifications as a spur to British industrialization.

Postlethwayt's *Universal Dictionary* (4th edn, 1774) purported to be a translation of Jacques Savary's *Dictionnaire universal du commerce*, but as Schumpeter (1954, pp. 156–7) noted, it was really much more. Nevertheless Schumpeter (p. 372, n.15) viewed Postlethwayt as a writer whose name survived despite 'substandard performance'. Schumpeter added that E.A.J. Johnson's careful bibliographic efforts 'reduced to its proper proportions the charge of plagiarism that has been frequently leveled against Postlethwayt, though the case remains bad enough' (Schumpeter, 1954, pp. 156–7). But Johnson himself concluded that his efforts 'relieve[d] Postlethwayt, at least partially, from an ill-founded charge' (Johnson, 1937, p. 405). Nonetheless, substantial portions of Richard Cantillon's *Essai* first appeared in English in Postlethwayt's *Dictionary* (Higgs, 1905, pp. ix–xiii) without acknowledgement.

Postlethwayt apparently sought, with mixed results, to become a well-heeled sycophant to British royalty through his work (Johnson, 1937, pp. 186–7). Johnson even speculated that Postlethwayt may have been a paid agent of the Royal African Company. He died abruptly in relative poverty in 1767 and was buried in Old Street churchyard in the Clerkenwell section of London. It is probable that he was the brother of James Postlethwayt, author of a major history of British public revenue.

Selected Works

1757a. *Britain's commercial interest explained improved*, 2 vols. New York: Augustus M. Kelley, 1968.

1757b. *Great Britain's true system*. New York: Augustus M. Kelley, 1967.

1774. *The Universal dictionary of trade and commerce*, 4th edn. New York: Augustus M. Kelley, 1971.

1968. *Selected works*, vol. 1, *1745–1757;* vol. 2, *1746–1759*. Farnborough/Hants: Gregg International Publishers.

References

Dorfman, J. 1971. Postlethwayt's pioneer British Commercial Dictionary. Preface to M. Postlethwayt, *The Universal dictionary of trade and commerce*. New York: Augustus M. Kelley.

Fay, C.R. 1934. *Imperial economy and its place in the formation of economic doctrine 1600–1932*. Oxford: Clarendon Press.

Higgs, H. 1905. Preface to W.S. Jevons, *The principles of economics: A fragment of treatise on the industrial mechanism of society and other papers*. London: Macmillan.

Johnson, E.A.J. 1937. *Predecessors of Adam Smith: The growth of British economic thought*. New York: Prentice-Hall.

Malachy Postlethwayt. In *Dictionary of national biography*, ed. L. Stephen and S. Lee. London: Oxford University Press, Vol. 16. Reprinted 1949–50.

Schumpeter, J.A. 1954. *History of economic analysis*. New York: Oxford University Press.

Viner, J.A. 1937. *Studies in the theory of international trade*. New York: Harper and Brothers.

Williams, E. 1944. *Capitalism and slavery*. Chapel Hill: University of North Carolina Press.

Postmodernism

M. Klaes

Abstract

Postmodernism resists encyclopaedic definition. On the level of economic phenomena, debates have centred on postmodernity as a separate historiographic period. On the conceptual level, the work of prominent economists has been argued to resonate with postmodernist themes. Certain parts of behavioural and experimental economics have begun to display key postmodernist features. A small selfconsciously postmodernist literature draws from economics, literary criticism, and Continental philosophical traditions in its analysis of economic phenomena.

Keywords

Animal spirits; Choice theory; Computational functionalism; Critical rationalism; Economic modernism; Enlightenment; Experimental economics; Keynes, J. M.; Postmodernism; Rhetoric of economics; Becker, G. S.; Rhetoric of economics; Methodology of economics; Relativism

JEL Classifications

B4

Postmodernism is a concept that escapes encyclopedic definition, to the extent that mischievous commentators have described postmodernists as a club of individuals who tacitly collude in a refusal to collectively define what postmodernism is about. This should strike a chord with any economists: who have been accused of leaving central notions such as market, firm, competition or equilibrium ill-defined (for example, Coase 1937; Clower 1995), despite having good grounds for doing so (compare Popper 1945: p. 18).

On the level of economic phenomena, debates have centred on whether or not one can consistently speak of postmodernity as a separate historiographic period. Advocates of postmodernity in this epochal sense assume that profound changes in the constitution of contemporary society have brought an end to the modern period, the close of which has variously been located from the last quarter of the 19th century to the last quarter of the 20th century. On the conceptual level, it has been argued that the work of several prominent economists, including Keynes (1936) and Becker (1976) for example, resonates with postmodernist themes. Broader strands of research in economics have begun to display key postmodernist features, most notably as a result of critical examination of the notion of the rationally unified individual. A small, selfconsciously postmodernist literature draws from economics, literary criticism, and Continental philosophical traditions in its analysis of economic phenomena.

Postmodernity

The postmodern found its initial motivation in postmodernity viewed as a historiographic category, commonly attributed to Arnold Toynbee (for example 1954, pp. 234–8). Toynbee suggested that the modern period in Western history, as the period immediately following the Middle Ages, had come to an end by the 1870s. He associated modernity with social stability, Enlightenment rationalism and progress. A 'post-Modern' period in turn was characterized by social unrest and the collapse of rationalism. This cultural pessimism in regard to the advent of the postmodern propagated by Toynbee and others contrasts with positive assessments of the move from industrialization to a post-industrial knowledge economy, where new technologies would replace ideology as key drivers of social change (for example, Toffler 1970). Both the culturally pessimistic and optimistic views share the acceptance of postmodernity as a particular historical phase with a distinct set of postmodern or 'late capitalist' socioeconomic features, a perspective which has found its apex in neo-Marxist stage theories of capitalist development (Mandel 1975; Jameson 1991).

The postmodern as a historiographic category rests on an epochal interpretation of history, which assumes that historic junctures separate adjacent periods. Many historians are not prepared to accept that modernity has been superseded by a qualitatively different period however. Interpreting the postmodern as 'post modernity' suffers here from the limitations that plague epochal categorization in general. To the extent that sceptics of postmodernity are not in fact sceptics regarding epochal categories and historiographies, they face a dilemma. They can either argue that modernity is the end of history (Fukuyama 1992), or propose an alternative successor to it. But how can one conceive of such alternatives as anything else than a particular interpretation of 'post modernity'?

Postmodern authors seek to avoid being trapped in binary oppositions of this kind. The work of Jean-François Lyotard, for example, has served as a prominent point of reference. His *Postmodern Condition* (1979) defined modernity in terms of a style of thought or epistemological outlook characterized by grand 'meta-narratives' centred on the ideas of scientific progress and individual emancipation, or the rationalist Enlightenment project *tout court*. Inverting these characteristics, Lyotard associated the postmodern with fragmented personal identities and a pervasive heterogeneity and indeterminacy of knowledge. But by doing so, he in fact affirmed the ahistorical dimensions of an ultimately bimodal categorization of contemporary society. The 'postmodern' turns thus into the less well-recognized face of the modern: '[p]ostmodernity is not a new age, but the rewriting of some of the features claimed by modernity . . . [T]hat rewriting has been at work, for a long time now, in modernity itself' (Lyotard 1987, p. 34).

Postmodernist Economics

In contrast to other social sciences, the notion of a postmodernist kind of economics has only relatively recently begun to gain currency, both as a label for the work of several prominent economists, including Keynes (1936) and Becker (1976) for example, and in terms of methodological features displayed in several strands of current research. To speak of postmodernist economics along these lines requires a concept of economic modernism to begin with. Largely unrecognized, two different understandings of economic modernism have sprung up, with different implications for the understanding of postmodernism in economics.

Economic modernism has been understood either as the manifestation in economics of modernism more generally understood as a widely recognized 20th-century socio-cultural style, or as the methodological face of modernity epochally conceived (see above). As a socio-cultural style, modernism is commonly thought to have flourished in the early 20th century, although, depending on the particular context, its influence may be traced from the late 19th century to the first decade of the 21st century and quite likely beyond it (compare Weston 1996). Across fields as diverse as literature, painting, music, architecture and design, proponents of modernism have questioned individual identity, displayed profound scepticism towards realist accounts of the world, and embraced dissonance and uncertainty as defining aspects of social life, developing ever more sophisticated forms of representation and a display of formal technique. What the many guises of modernism share is a profound reaction to the conditions of modernity.

In contrast to this avant-garde notion of modernism as the pursuit and transcendence of the limits of modernity, the concept of modernism first entered economics in a more restricted and conservative way, encapsulating a rejection of the methodological face of modernity. Economists by and large see themselves as adhering to the broad outlines of a critical rationalist methodology. This 'official' methodology of economics has been characterized by some methodologists as 'modernist', in the sense that it is committed to a belief in scientific progress through the formulation and empirical testing of hypotheses, to the rational actor paradigm, and to mathematical formalism (McCloskey 1983; Dow 1991).

Most economists will, of course, find these assumptions innocuous. It is thus no coincidence that economic modernism, in the sense described, is typically employed by authors who dissent from the 20th-century neoclassical tradition in economics. Samuelson's (1939) article on the multiplier–accelerator model, itself a central contribution to Keynesian business cycle theory, has been cited as a *pièce de résistance* in this regard, illustrating the modernist spirit underlying the neoclassical school (Klamer 1995). The article covers barely four journal pages. Packed with mathematical notation, tables and graphs it keeps discursive elements to a minimum. Compared with Samuelson's treatment of the business cycle, Keynes's (1936) original analysis engages in an exuberance of narrative in his explanation of the business cycle, coming to a head in the well-known passages of Chap. 22 of the *General Theory.* Keynes's portrayal there of the uncertainties of the world of markets as being beyond the reach of rational analysis and containable only within a domain of 'animal spirits' (though channelled by social conventions) has prompted some authors to regard these aspects of his work as indicative of important postmodernist currents in 20th-century economics (in particular Ruccio and Amariglio 2003, Chap. 2), which reflect concerns comparable to the appreciation of the heterogeneity and indeterminacy of knowledge as it can be found in the work of Lyotard (1979), for example.

Reading the work of the mature Keynes as an expression of postmodern currents in the economics of the 1930s requires us to regard the neoclassical orthodoxy of that time as the prime manifestation of economic modernism, thereby allowing a rapprochement between dissenting schools of thought and a postmodern kind of economics. Not all who have identified postmodern aspects in economics have found compelling the association of economic postmodernism with dissenting, and of economic modernism with consenting, approaches vis-à-vis a putative mainstream, however. Characterizing Keynes's work as postmodernist can be challenged on historiographic grounds along the lines discussed above in the context of epochal interpretations of the postmodern. Moreover, this characterization rests on interpreting economic modernism as the methodological face of modernity. If modernism is instead understood as a broadly based early 20th-century socio-cultural style, there are good grounds for regarding the *General Theory* and other works of Keynes as a prime expression of economic modernism (Klaes 2006).

Rather than depicting it as a caricature of orthodox approaches in economics, the appreciation of an economic modernism in its own right may help to account for a range of departures from the neoclassical tradition in early 20th-century economics, including both Keynes's *General Theory* and the work of Samuelson and others who were at the forefront of the formalist turn in economics. Conversely, postmodernist dimensions in economics may be sought not only in dissenting approaches, a point most prominently expressed by Jameson (1991, pp. 263–71), who argues that Becker's (1976) work, in its treatment of children, companionship and health as conventional commodities, displays a deep affinity with the postmodernist notion of consumption as an all pervasive cultural pattern, sharing the ambition of reducing all human interaction to market exchange.

Decentred Economic Selves

As an illustration of how close contemporary theorizing in economics has come to key postmodernist concerns, let us consider how individuals are portrayed in economics. According to Davis (2003), there no longer exists a coherent account of the individual in contemporary economics following its de-psychologization and reduction to a rational preference ordering. With no concept of the individual beyond this ordering, choice theory has become equally applicable to individual persons and supra-person individuals like firms. Increasingly, however, economists also entertain the possibility of multiple sub-person objectives, with fascinating challenges to the notion of a unified economic self (for example, Schelling 1984).

Ever since the publication of Berle and Means (1932), economists have been attuned to the split personality of multi-person individuals. Senior management follow their own objectives that do not necessarily coincide with the objective function of the corporation. A similar line of argument can be applied to the notion of a coherent self. At the sub-person level we may well consist of a range of competing selves. The de-psychologization of the individual in economics leads therefore to a postmodern critique of integrated individual identities.

Economists have begun exploring the implications of a decentred economic self (Kavka 1991; Steedman and Krause 1986), which rests on the proposition that the market without is matched by a market within. Warding off a postmodernist disintegration of the unified self amounts to solving the internal 'social choice' problem through the imposition of a dictator. To the extent that the internal and external worlds of choice are formally equivalent however, this literature has revealed a curious asymmetry whereby the desirability of dictatorial solutions to the internal choice problem is taken for granted in the same vein as its undesirability regarding the external choice problem is taken for granted in economics.

Upon closer inspection, choice theory exhibits further postmodernist dimensions. Pursuing his basic argument from another angle, Davis (2003) suggests that economics, in its rejection of early neoclassical subjectivism, has subscribed to computational functionalism in its conception of the abstract individual. Computational functionalism, as a theory of mind, holds that brain states are computational states of mental algorithms, and that two individuals share the same type of mental state if they function in a causally equivalent way in respect of their physical environment. The abstract individual is therefore a preference computing algorithm, boundedly rational or not, that can be implemented in different entities without prejudice as to whether these entities are individual human beings, particular 'modules' within a human brain, economic institutions such as firms and markets, or non-humans (animals, machines and other 'aliens').

Mirowski (2002) has cast this ontological indifference of the economic individual in respect of its range of potential actualizations (human decision-makers, various subsets of brain tissue, animals, computers, and so forth) into the postmodern motif of the cyborg, a cyber organism that is half man and half machine. Recent work in experimental economics has unwittingly come up with an interesting illustration of this proximity between man and machine: on the level of convergence and efficiency, double auction behaviour of experimental subjects and computational agents programmed as random number generators turns out to be rather similar (Gode and Sunder 1993). The cognitive capacities of market participants matter much less than the market algorithm itself. This allows the provocative suggestion that markets use us simply as pawns to further their algorithmic life, with the possibility of endogenous evolution of cyborg-like market automata in a decidedly post-humanist and thereby postmodernist fashion.

While postmodernist dimensions become apparent in a range of current strands of research in economics once they are read with an eye sensitive to debates in other social sciences and the humanities, self-consciously postmodernist work in economics has remained relegated to its fringes (see the collections by Woodmansee and Osteen 1999; Cullenberg et al. 2001; and Zein-Elabdin and Charusheela 2004). Its most influential impetus has come from the rhetoric of economics tradition. Initially concerned with the rhetoric found in the texts of academic economists, rhetoricians of economics have generalized their approach to include economic conversation more generally conceived (McCloskey 1994, pp. 367–78). Prices are carriers of information only because they are part of a conversation. Entrepreneurs succeed only if they can persuade others to provide the capital necessary for turning their inventions into marketable products, which only sell if consumers can be persuaded to buy them. Stock markets epitomize this conversational feature of economic life (see Shiller 1989, p. 56, 387).

The resulting suggestion, of reading the economy as a text (Brown 1994), leads to the prospect of a postmodernist economics that approaches the economy on the premise that all economic texts should be treated alike in this ongoing overarching conversation by which resources are allocated, be they authored by Nobel Laureates or the man or woman on the street. Opinion is divided even among rhetoricians regarding the merit of so radical a revision of the traditional hierarchy between economic analyst and agent (Mehta 1999; McCloskey 1999). It points to the implied relativism present in most postmodernist perspectives as a major and recurring point of contention (see Backhouse 1998), although relativism as such, though not popular among practitioners and methodologists alike, is not the unanimously discredited philosophical position that some make it out to be (Kusch 2002).

See Also

- ▶ Culture and Economics
- ▶ Economic History
- ▶ Methodology of Economics
- ▶ Pluralism in Economics
- ▶ Rhetoric of economics

Bibliography

Backhouse, R.E. 1998. Should economists embrace postmodernism? In *Explorations in economic methodology. From Lakatos to empirical philosophy of science*, ed. R.E. Backhouse. London: Routledge.

Becker, G.S. 1976. *The economic approach to human behavior*. Chicago: University of Chicago Press.

Berle, A.A., and G.C. Means. 1932. *The modern corporation and private property*. New York: Legal Classics Library.

Brown, V. 1994. The economy as a text. In *New directions in economic methodology*, ed. R.E. Backhouse. London: Routledge.

Clower, R.W. 1995. Axiomatics in economics. *Southern Economic Journal* 62: 307–319.

Coase, R.H. 1937. The nature of the firm. *Economica, New Series* 4: 386–405.

Cullenberg, S., J. Amariglio, and D.F. Ruccio. 2001. *Postmodernism, economics and knowledge*. London: Routledge.

Davis, J.B. 2003. *The theory of the individual in economics*. London: Routledge.

Dow, S. 1991. Are there any signs of postmodernism with economics? *Methodus* 3: 81–85.

Fukuyama, F. 1992. *The end of history and the last Man*. London: Hamilton.

Gode, D., and S. Sunder. 1993. Allocative efficiency of markets with zero- intelligence traders. *Journal of Political Economy* 101: 119–137.

Jameson, F. 1991. *Postmodernism: Or, the cultural logic of late capitalism*. Durham: Duke University Press.

Kavka, G.S. 1991. Is individual choice less problematic than collective choice? *Economics and Philosophy* 7: 143–165.

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

Klaes, M. 2006. Keynes between modernism and post modernism. In *The Cambridge companion to Keynes*, ed. R.E. Backhouse and B.W. Bateman. Cambridge: Cambridge University Press.

Klamer, A. 1995. *The conception of modernism in economics: Samuelson, Keynes and Harrod*. Aldershot: Elgar.

Kusch, M. 2002. *Knowledge by agreement*. Oxford: Clarendon.

Lyotard, J.-F. 1979. *La condition postmoderne*. Paris: Éditions de Minuit.

Lyotard, J.-F. 1987 [1991]. Rewriting modernity. In *The inhuman: Reflections on time*, ed. J.-F. Lyotard. Trans. G. Bennington and R. Bowlby. Cambridge: Polity.

Mandel, E. 1975. *Late capitalism,* Rev. edn. Trans. J. de Bres. London: Verso.

McCloskey, D.N. 1983. The rhetoric of economics. *Journal of Economic Literature* 21: 48–517.

McCloskey, D.N. 1994. *Knowledge and persuasion in economics*. Cambridge: Cambridge University Press.

McCloskey, D.N. 1999. Jack, David, and Judith looking at me looking at them. In *What do economists do? New economics of knowledge*, ed. R.F.J. Garnett. London: Routledge.

Mehta, J. 1999. Look at me look at you. In *What do economists do? New economics of knowledge*, ed. R.F.-J. Garnett. London: Routledge.

Mirowski, P. 2002. *Machine dreams. Economics becomes a cyborg science*. Cambridge: Cambridge University Press.

Popper, K.R. 1945. *The open society and its enemies*. London: Routledge.

Ruccio, D.F., and J. Amariglio. 2003. *Postmodern moments in modern economics*. Princeton: Princeton University Press.

Samuelson, P.A. 1939. Interactions between the multiplier analysis and the principle of acceleration. *Review of Economics and Statistics* 21: 75–78.

Schelling, T.C. 1984. Self-command in practice, in policy and in a theory of rational choice. *American Economic Review* 74: 1–11.

Shiller, R.J. 1989. *Market volatility*. Cambridge, MA: MIT Press.

Steedman, I., and U. Krause. 1986. Goethe's *Faust,* arrow's possibility theorem and the individual decision-taker. In *The multiple self*, ed. J. Elster. Cambridge: Cambridge University Press.

Toffler, A. 1970. *Future shock*. New York: Random House.

Toynbee, A. 1954. *A study of history*. Vol. 9. New York: Oxford University Press.

Weston, R. 1996. *Modernism*. New York: Phaidon.

Woodmansee, M., and M. Osteen. 1999. *The new economic criticism. Studies at the intersection of literature and economics*. London: Routledge.

Zein-Elabdin, E.O., and S. Charusheela. 2004. *Postcolonialism meets economics*. London: Routledge.

Poverty

A. B. Atkinson

JEL Classifications
D6

Concern for poverty has been expressed over the centuries, even if its priority on the agenda for political action has not always been high. Its different meanings and manifestations have been the subject of study by historians, sociologists and economists. Its causes have been identified in a wide variety of sources, ranging from deficiencies in the administration of income support to the injustice of the economic and social system. The relief, or abolition, of poverty has been sought in the reform of social security, in intervention in the labour market, and in major changes in the form of economic organization.

Poverty today is most obvious – and has the most pressing claim on our attention – on a world scale. The unequal distribution of income between countries, and the disparities within countries, mean that there are large numbers of people in Africa, Asia and Latin America whose standard of living would be agreed by everyone to be poor. The World Bank has suggested that there is 'a global total of close to 1 billion people living in absolute poverty' (World Bank 1982, p. 78), of whom about 400 million are thought live in South Asia, about 150 million in China, and some 100 million in East/South-East Asia and Sub-Saharan Africa. At such levels of living, the risks of death through hunger or cold, and vulnerability to disease, are of a quite different order from those in advanced countries. This has manifested itself most urgently in the occurrence of famine. Whatever the immediate cause of such disasters, whether inadequate total supply of food or whether unequal distribution, the severity of the situation in areas such as the Sahel and Ethiopia is an indicator of the precariousness of survival in many low income countries.

Such mass poverty in poor countries is quite different from poverty in advanced countries. The target of the American War on Poverty, launched in 1964, was the minority of Americans with incomes below a poverty line of $3000 a year for a family of four (in 1962 prices), which was many times the average income of India. The basis for the US official poverty line is to be found in a food consumption standard (the Department of Agriculture economy food plan), but its level reflects the prevailing living conditions in that society. It might well be argued that concern with poverty in advanced countries, at a time when other countries face disaster, is unjustified and that the term 'poverty' cannot legitimately be applied. The parallel may be drawn with rearranging the deckchairs on the *Titanic* as the ship goes down. This does not, however, seem fully apposite. A closer parallel is with the position of those on ships steaming to the aid of the stricken vessel. The overriding objective should be to get to the rescue as rapidly as possible, but those on the rescuing ships should also be concerned that their steerage passengers do not die of exposure on the way. The relief of famine, and the redistribution of income to alleviate poverty on a world scale, should have priority, but the problem of poverty in advanced countries, defined in their terms, may legitimately come next on the list of concerns.

The fact that the term 'poverty' is being used in different senses highlights the need to clarify the underlying concept and the discussion so far has touched on several aspects which need to be elaborated. After a brief historical review of studies of poverty in section "Historical Review of Studies of Poverty", we examine some key conceptual issues. What is the indicator of resources which should be employed in measuring poverty? What is the underlying notion of poverty and how is it related to inequality? These issues are discussed in section "Poverty: Living Standards and Rights". The determination of the poverty standard is a crucial question. Here we need to consider approaches based on such 'absolute' concepts as food requirements and those poverty scales which are explicitly) 'relative'. We must consider the treatment of families with differing needs. These topics are the subject of section "Setting the Poverty Line". Once we have established the extent of poverty, its causes become a central concern. Here we are led first to ask 'who are the poor?' This is examined in section "The Composition of the Poor". Is poverty concentrated in particular classes or particular sections of society? How far is it associated with particular stages of the life-cycle? The composition of the poor provides in turn a starting point for the investigation of the underlying causes of poverty, and an analysis of policies to combat poverty. These are the subject of section "Causes and Policies".

Historical Review of Studies of Poverty

The scientific study of poverty in the Anglo-Saxon world is usually taken to date from the investigations of Booth and Rowntree at the end of the 19th century. In Britain it is true that King and others had given estimates of the number of paupers; and that *The State of the Poor* by Eden (1797) contained a great deal of material collected from over 100 parishes and giving details of family budgets. Engels and Mayhew provided insight into the condition of the poor in urban England. But it was Booth's *Life and Labour* (1892–7) survey of London, started in the East End in the 1880s, that combined the elements of first-hand observation with a systematic attempt to measure the extent of the problem. Taking the street as his unit of analysis, he drew up his celebrated map of poverty in London.

The study of Rowntree (1901) was intended to compare the situation in York, as a typical provincial town, with that found by Booth in London, but his method represented a significant departure in that it was concerned with individual family incomes and in that he developed a poverty standard based on estimates of nutritional and other requirements. The development of survey methods was taken further by Bowley (1912–13) who pioneered the use of sampling in his 1 in 20 random sample of working-class households in Reading. A great many local studies were subsequently conducted, including Bowley's Five Towns survey in 1915, replicated in the early 1920s, and the new Survey of London Life and Labour published in the early 1930s. Rowntree himself repeated his survey of York in 1936 and 1950. The latter became the standard source of information as to the effectiveness of the post-1948 welfare state, with most commentators concluding that poverty had been effectively abolished in Britain by the combination of full employment and the new social benefits. Doubt began to be cast on this conclusion by the work of empirical sociologists and came to the fore with the publication of *The Poor and the Poorest* by Abel-Smith and Townsend (1965). This showed, using secondary analysis of a national survey, that in 1960 about two million people fell below the social security safety net level. This finding was confirmed in official estimates which began to be published by the Department of Health and Social Security in the 1970s, and by Townsend's own major survey (1979).

As in many fields, the United States entered later and has taken the subject further. The definition of a poverty line was attempted by Hunter in 1904 and this was developed in a series of studies, such as the 'minimum comfort' and other budgets produced for New York City. There was the 1949 report on low income families by the Joint Committee on the Economic Report. It was not however until the 1960s that the problem of poverty received systematic study, with a few notable exceptions such as the work of Lampman (1959). *The Other America* by Harrington (1962) and *The Affluent Society* by Galbraith (1958) did much to arouse the attention of the public, politicians and academics. The 1964 report of the Council of Economic Advisers set out the $3000 poverty level, drawing heavily on the research of Orshansky (1965), and this was subsequently refined to form the official poverty line, which has been applied since that date (with modifications, such as the addition of alternative measures including the value of transfers in kind).

Similar studies have been carried out in many countries, and researchers have become increasingly interested in cross-country comparisons. The OECD made an early attempt at such comparisons and a more extensive exercise is being carried out in the Luxembourg Income Study. Any assessment of world poverty depends on the availability of information about the distribution of living standards within individual countries; and here both the World Bank and the International Labour Organization have made significant contributions. In some low income countries, there has been extensive research on poverty, India being an example, where there has been a great deal of discussion as to whether poverty has increased or decreased over time. The ILO and the World Bank have also been influential in the widespread interest, reflected in the Brandt Report (1980), in the concept of 'basic needs', or a minimum set of specific goods and environmental conditions.

Poverty: Living Standards and Rights

Concern about poverty may take the form of concern about such basic needs: for example, food, housing and clothing. In this case, we can identify clearly the items of consumption in which we are interested. This approach leads to poverty being measured in a multidimensional way, where a family may be deprived in one but not other respects, although particularly serious will be situations where families suffer deprivation in several dimensions, or what is referred to typically as 'multiple deprivation'.

This approach is concerned with specific deprivation, but we may also seek to record disadvantage in a single index of living standards, such as total expenditure, a household being said to be in poverty if it has total expenditure below a specified amount. This is not however the approach followed in most studies of poverty in advanced countries, which record poverty on the basis of total *income*. Income may *understate* the level of living. A family may be able to dissave or to borrow, in which case its current level of living is not constrained by current income and expenditure may be the more appropriate index. (Although in the short run there may be a divergence between *consumption* and *expenditure*, as families use up stocks of goods, etc.) The level of living may exceed that permitted by income where the family is able to share in the consumption of others. An elderly person living with his or her children may benefit from their expenditure. Income may, conversely, *overstate* the level of living. This may happen where money alone is not sufficient to buy the necessary goods: where there is rationing, or unavailability of goods. It is also possible that people choose a low level of consumption. This latter reason has led to its being argued that income *should* be the indicator of poverty, since it is a measure of the opportunities open to a family and is not influenced by the consumption decisions made.

In considering the choice between income and expenditure, it is helpful to distinguish two rather different conceptions of poverty: that concerned with *standards of living* and that concerned with *minimum rights* to resources. On the former approach, the goal is that people attain a specified level of consumption (or consumption of specific goods); on the latter approach, people are seen as entitled, as citizens, to a minimum income, the disposal of which is a matter for them. In practice, the two notions are often confounded, but the distinction is important, and it has obvious implications for the choice of poverty indicator. Income is the focus of the rights approach, but its use on a standard of living approach must be seen as a proxy for consumption.

The reference to 'rights' raises the question of the relation between poverty and inequality. Here four different schools of thought may be distinguished. There are those who are concerned only with poverty, attaching no weight to income inequalities above the poverty line. There are those who attach weight to the reduction of inequality as a goal of policy but give priority to the elimination of poverty, so that we have a lexicographic objective function. There are those who are concerned about both goals and who are willing to trade gains in one direction against losses in the other. Finally, there are those who attach no especial significance to poverty, simply regarding it as a component of the wider cost of inequality.

In this context, reference should be made to the choice of *poverty measures.* Where poverty puts survival in doubt, it is natural to take as one's measure the proportion of the population at risk. Concern for minimum rights may also make the 'head count' the most relevant measure. But we may also be concerned, particularly on a standard of living approach, with the severity of poverty, in which case measures such as the poverty deficit (the total shortfall from the poverty line) may be more appropriate. One can indeed go further, as proposed by Sen (1976), and take account of the distribution of income within the poor population: for example, with the poverty index depending on the Gini coefficient for this distribution.

Setting the Poverty Line

The most straightforward approach to the determination of the poverty line is to specify a basket of goods, denoted by the vector $\mathbf{x}^*$, purchasable at prices $\mathbf{p}$, and to set the poverty standard as:

$$(1+h)\mathbf{p}\cdot\mathbf{x}^*$$

where h is a provision for inefficient expenditure or waste, or a provision for items not included in the list $\mathbf{x}$*. This was in effect the method adopted by Rowntree, whose diet for Tuesdays was porridge for breakfast, bread and cheese for lunch, and vegetable broth for dinner. It was the method followed by Orshansky, where $\mathbf{x}^*$ represented food requirements and h (=2) made allowance for spending on other goods. This approach is often referred to as an 'absolute' poverty standard, and contrasted with a 'relative' approach that relates the poverty line to contemporary levels of living: for example the proposal of Fuchs in the United States that the poverty line should be one-half the median family income. It is sometimes suggested that the absolute standard is less problematic than the relative approach and less dependant on value judgements.

The term 'absolute' can, however, scarcely be used in the same sense as in the physical sciences and there is scope for a great deal of disagreement about where the line should be drawn. This is most evident in the case of the rights approach, where the determination of the minimum level of income is explicitly a social judgement, but it applies also to the standard of living approach. In the case of food requirements, where a physiological basis may appear to provide a firm starting point, it is in fact difficult to determine $\mathbf{x}^*$ with any precision. There is no one level of food intake required to survive, but rather a broad range where physical efficiency declines with a falling intake of calories and protein. Nutritional needs depend on where people live and on what they are doing. They vary from person to person, so that any statement can only be probabilistic: at a certain level of consumption there is a certain probability that the person is inadequately fed. Even if these problems could be resolved, there is the difficulty of the disparity between expert recommendations and actual consumption behaviour. The factor h is intended to allow for this, but the precise allowance will depend on the judgement of the investigator. Rowntree, for example, included an allowance for tea, which has little or no nutritional value but which formed a staple item of consumption.

In the case of non-food items, there is even greater scope for judgement. This applies whether we seek to include the goods in the vector $\mathbf{x}^*$ or whether we allow for non-food items via the multiplier h. For example, the procedure of Orshansky has been criticized as under-stating the proportion of income spent on food and hence overstating the value of h. More fundamentally, the role of goods in the determination of the poverty line needs reconsideration. The literature on 'household production' has pointed to the role of goods as an input into household activities, with the level of activities being our main concern rather than the purchase of goods as such. On this basis, if we denote the target level of activities by $\mathbf{z}^*$, and if there is an input–output matrix A, relating goods inputs to activity levels, then the necessary level of expenditure becomes:

$$Y=(1+h)\mathbf{p}A\mathbf{z}^*$$

The significance of this view is that poverty may be measured in absolute terms, in the sense that the vector $\mathbf{z}^*$is fixed, but the required bundle of goods may be changing because the input-output matrix is affected by developments in the particular society. If the activity is 'attending school', then the demands in terms of clothing, books and equipment are quite different today from those of a century ago. This does not mean that there is no distinction between absolute and relative concepts. There is a clear difference in principle between taking the vector $\mathbf{z}^*$ as fixed and allowing it to be influenced by the living patterns of the rest of society, as in the work of Townsend (1979), who is concerned with the extent to which families can participate in the 'community's style of living'.

The notion of a fixed absolute poverty standard, applicable to all societies and at all times, is therefore a chimera. Nor is it evident that a poverty standard, once set, can be compared across time by simply adjusting by an index of consumer prices. In the case of both absolute and relative approaches, we have to face the problems of judgement. Here several lines of attack may be discerned. There are studies which take the *official* poverty standards as embodying social values, which seems natural on the minimum rights approach and which at least provides a measure of governmental performance. There are studies which base the poverty line on the views expressed in surveys of the population as a whole. In the United States, the Gallup Poll has regularly asked the question: 'What is the smallest amount of money a family of four needs to get along in this community?' These, and other approaches, will produce a range of poverty lines, and it seems unlikely that we can reach universal agreement. There are therefore strong reasons for recognizing such differences of view explicitly and using a *range* of poverty lines. This means that we may not be able to reach unambiguous conclusions – it may be that poverty will be shown to have increased according to one line but not according to another – but it will avoid a total impasse. In the same way, when making a comparison over time, we may want to compare 1950 with two alternative lines for 1980, one updated by the price index and the other adjusted to allow for rising real incomes, thus generating a 'confidence interval' around the 1980 estimate.

To this point, the poverty line has been discussed as though it were a single number, but families of different types and different sizes will receive different treatment. In Britain, for example, the social security safety net is typically some 60 per cent higher for a couple than for a single person. The relationship between the poverty lines for different family types is usually referred to as an *equivalence scale.* However, a prior question before the equivalence scales are determined is the choice of the *unit of analysis.* Here the distinction between the standard of living and rights approaches is important. In the latter case, the notion of rights must be essentially individualistic. The case for considering a wider unit must rest on there being within- family transfers which cannot be adequately observed. The family is taken when measuring poverty because we do not accept that a large number of those with zero recorded cash income are in fact without resources. At the same time, little is known about the distribution of income within the family. Certainly, it would be quite wrong to treat all married couples as having equal rights to the joint income. On a standard of living approach, the logical unit is that which shares consumption; and we may wish to go beyond the inner family to the household as a whole. This would take account of the fact that items of expenditure may have 'public good' characteristics for the family members. Again, however, it may be that there are unequal living standards within the household.

Several approaches have been adopted to the determination of the equivalence scales for different-sized units. Survey information about individual assessments of what is needed 'to get along' has been used for this purpose. More commonly, the basis has been sought for observation of actual behaviour. One of the early methods provides an illustration. By taking a commodity consumed only by adults (e.g. men's clothing), one can observe the level of income at which a family with one child, say, can attain the same level of consumption of that commodity as a family with no children. This method, and other more sophisticated implementations of the idea, have been the subject of considerable debate. The underlying difficulty is that one is assuming, in the example given, that preferences for the commodity are independent of family composition: the arrival of the child may mean that the couple go out less and spend less on clothing. With other methods based on observed consumption behaviour, identifying restrictions are similarly needed. At a more fundamental level, the ethical status of such scales is far from transparent. Not only is it impossible to draw conclusions about welfare levels with different family compositions, but also society may wish to modify the implied judgements: for example, to vary the parental evaluation to take account of the interests of the children.

The Composition of the Poor

One of the main aims of those investigating poverty has been to establish who the poor are. Popular opinion is often coloured by vivid, but not necessarily representative, accounts of life below the poverty line. For this reason, the Council of Economic Advisers stressed at the start of the War on Poverty in the US that poverty should not be seen as a minority phenomenon: 'Some believe that most of the poor are found in the slums of the central city, while others believe that they are concentrated in areas of rural blight. Some have been impressed by poverty among the elderly, while others are convinced that it is primarily a problem of minority racial and ethnic groups. But objective evidence indicates that poverty is pervasive. . .the poor are found among all major groups in the population and in all parts of the country' (1964, pp. 61–2).

Poverty in advanced countries affects a minority in terms of numbers but it is not confined to specific marginal groups. At the same time, certain groups are much more at risk. In 1983, the poverty rate for blacks in the United States was nearly three times that for whites, and that for Hispanics was more than twice. Compared with the average, the rate for families with children is nearly double, and that for families with a female head is much higher. The evidence for other countries equally shows large differences in the incidence of poverty between groups: for instance, in Malaysia, recorded poverty among Malays is much higher than among the Indian or Chinese ethnic groups. The World Bank has argued that poverty in low income countries is very much a rural problem; and the evidence from India shows poverty to be much higher in rural than urban areas.

If we seek to probe further into the composition of the poor, then the dynamics of poverty must be taken into account. Is poverty a largely transitory phenomenon, in that the families poor today will quite probably be above the poverty line next year? Is poverty associated with particular periods of the life cycle? Transitory poverty may occur for a variety of reasons. Income may be temporarily reduced because of ill-health or unemployment or because wages are cut. It may be a bad harvest. Families may split up, leaving one parent with the family responsibilities but inadequate income. The evidence from panel surveys, where the same families are interviewed on a continuing basis (as, for example, in the Michigan Panel Study of Income Dynamics), has shown the extent of mobility in the incomes and circumstances of the poor. A sizeable fraction of those recorded as poor in one year are above the poverty line next year. This does not mean that their poverty is not a matter for concern, since low current incomes may impose severe hardship, but it means that these people do not constitute a permanent 'under class'.

Such mobility does however require careful interpretation. It may arise on account of the life cycle. In Rowntree's 1899 survey he found that the life of the labourer was marked by 'five alternating periods of want and comparative plenty', the periods of want being childhood, when he himself had children, and old age. The impact of such life-cycle factors depends on the extent to which income support is provided by state or private transfers. In this respect the situation in Britain has changed dramatically since 1899, with the introduction of state pensions, a large increase in private pensions, and the payment of child and other benefits. In other countries too there has been major growth in transfers: between 1960 and 1981 social expenditure as a percentage of GDP rose in the United States from 7 per cent to 15 per cent, in West Germany from 18 per cent to 27 per cent, and in Japan from 4 to 14 per cent (Institute for Research on Poverty 1985). Transfers, and other programmes, such as health care, must have reduced the extent of life-cycle poverty. The incomes of the elderly in the United States, for example, are considered to have risen relative to those of the population as a whole. But there remains concern about certain stages of the life cycle, particularly among families with children; and while the poverty rate among the elderly in the US has fallen, that among the non-elderly has risen.

To the extent that poverty is a life-cycle phenomenon, this means that more people experience poverty at some point in their lives but that its duration is limited. At the same time, poverty at one stage of the life cycle may lead to poverty at a subsequent stage. Those who are hard-pressed when they are bringing up children may have little savings on which to draw in retirement. Those who grow up in low- income families may themselves be more likely to be below the poverty line, as was found in the follow-up in the 1970s of the children of the families interviewed by Rowntree in 1950 (Atkinson et al. 1983). Moreover, we should not however lose sight of the fact that for some people poverty persists. Agricultural labourers, or farmers with small plots, may be in poverty even in 'good' years. Among industrial workers, there are those whose earnings are inadequate to support even themselves; there may be a problem of low *pay.* And the low paid may be more vulnerable to the transitory factors such as ill-health and unemployment.

Causes and Policies

In 1913, R.H. Tawney argued for the restatement of the problem of poverty: 'the diversion to questions of social organization of much of the attention which, a generation ago, was spent on relief'. The problem of poverty, he said, was 'primarily an industrial one'. In terms of the composition of the poor described above, this means that the causes of poverty were sought not in the failure of income support but in the reasons why income was inadequate in the first place.

Tawney recognized the importance of personal factors in causing poverty, but laid principal stress on the position of groups and classes and their economic situation, factors which may equally be relevant today. Workers may be locked into low-paying industries where techniques and machinery need to be modernized; they may live in depressed regions to which private capital cannot be attracted. There may be a low level of unionization and employers may be able to hold wages down.

These aspects, which have been emphasized in theories of 'segmentation' in the labour market, point to the need for government intervention. This may take the form of minimum wage legislation, to guarantee minimum levels of earnings, coupled with measures to offset any adverse effect on employment and to modernize the sectors or regions concerned. At a macro-economic level, the government has an important responsibility. Studies in the United States have identified unemployment as a much more serious problem than inflation for low income groups. There can be little doubt, for example, that the recession of the 1980s has increased the incidence of poverty in advanced countries.

The counterpart of this structural explanation in the context of less developed, primarily agricultural economies is to be found in the role of land tenure and its distribution and in the nature of labour and capital markets. Rural poverty is high among landless labourers and those farmers with small or unproductive holdings. Their difficulties may be intensified by the terms on which they have to borrow or purchase intermediate goods. Here too policy requires government intervention, whether to redistribute land holdings, or to facilitate the introduction of new methods, or to eliminate extortionate lending practices, or to provide non-farm employment. Measures such as land reform raise major political issues, and in both developing and advanced countries it can be argued that basic changes in the form of economic system are necessary to eradicate poverty. The World Bank has noted, for example, the role played by the Chinese food security policy in the reduction of poverty and the way in which it is tied into China's collective system.

The industrial explanation of poverty may be contrasted with the 'supply side' explanation which has seen low pay as attributable to workers lacking productive skills, because they have been unable to complete education or training. This 'human capital' interpretation leads in turn to the policy recommendation that training and educational programmes should be expanded, a proposal that is congruent with the goal of reducing inequality of opportunity. Education and training had a central role in the United States War on Poverty, with schemes such as the Job Corps and the Neighborhood Youth Corps. A characteristic of individual workers also identified in the United States is that of race. Discrimination may lead to otherwise equally qualified workers receiving lower pay, as where black workers were prevented from entering certain occupations. The civil rights legislation and the operations of the Equal Employment Opportunity Commission may have reduced the direct effect of discrimination (as well as the indirect effect via unequal opportunities in education, etc.), but although the policy implications are clear in principle, experience suggests that they are not easily made effective.

Policies to improve job and earnings prospects must be central to the elimination of poverty, but they cannot succeed without complementary income maintenance provisions. The growth of transfers has not succeeded in providing a completely effective income guarantee for those without incomes from work or with additional needs. This is because of incomplete coverage, particularly where new needs develop, because of the inadequate levels of benefits (for example, those paid to people with poor employment records) and the incomplete take-up of income-tested benefits. In the last case, there is evidence that complexity or stigma deters families from claiming the transfers to which they are entitled, and hence they fall through the safety net.

To this end, proposals have been made for major reform of the transfer systems in advanced countries. One front-runner for many years in the United States has been the 'negative income tax', which would pay an income-related supplement using the income tax machinery. There are those reformers who would like to integrate fully the income tax and social security systems, as with the basic income guarantee scheme, where everyone receives a basic income and is then taxed on all income. Such a reform would mean that income maintenance largely ceased to be categorical: for example, there would not be separate treatment for the unemployed or the sick. An alternative would be to preserve the categorical nature of social insurance but to make the insurance benefits more extensive in their coverage and sufficient to avoid the necessity to depend on public assistance or other forms of means-tested benefits. In considering the feasibility of such reforms, one must have regard both to the arithmetic of the redistribution and to the reasons why they have not been enacted in the past. As the 'public choice' school of public finance economists has stressed, the actions of the government are themselves to be explained by economic and other motives. The reasons why governments have failed to enact successful anti-poverty policies is a subject of great importance.

The policies discussed in this section are solely concerned with the poverty *within* countries, and would do nothing to redistribute between countries. Indeed, some of the policies designed to help the low paid in advanced countries may actually have adverse consequences for low income countries. The income transfers which rich countries have so far made are of minuscule size when viewed against the magnitude of the problem of world poverty, and there can be little doubt that redistribution on a world scale is of the highest priority.

See Also

- ▶ Equality
- ▶ Poor Law, New
- ▶ Poor Law, old

Bibliography

Abel-Smith, B., and P. Townsend. 1965. *The poor and the poorest*. London: Bell.

Atkinson, A.B.., A.K. Maynard, and C.G. Trinder. 1983. *Parents and children*. London: Heinemann.

Booth, C. 1892–7. *Life and labour of the people of London*. 9 vols. London: Macmillan.

Bowley, A.L. 1913. Working class households in Reading. *Journal of the Royal Statistical Society* 76: 672–701.

Brandt, W. 1980. *North–South*. London: Pan.

Council of Economic Advisers. 1964. *Annual report*. Washington, DC: Government Printing Office.
Eden, F.M. Sir. 1797. *The state of the poor*. London: Cass.
Galbraith, J.K. 1958. *The affluent society*. Boston: Houghton Mafflin.
Harrington, M. 1962. *The other America*. New York: Macmillan.
Institute for Research on Poverty. 1985. Antipoverty policy: Past and future. *Focus*, Summer.
Lampman, R.J. 1959. The low income population and economic growth. Study paper no. 12, US Congress Joint Economic Committee. Washington, DC: Government Printing Office.
Orshansky, M. 1965. Who's who among the poor: A demographic view of poverty. *Social Security Bulletin* 28 (7): 3–32.
Rowntree, B.S. 1901. *Poverty*. London: Macmillan.
Sen, A.K. 1976. Poverty: An ordinal approach to measurement. *Econometrica* 44 (2): 219–231.
Tawney, R.H. 1913. *Poverty as an industrial problem*. London: London School of Economics.
Townsend, P. 1979. *Poverty in the United Kingdom*. London: Penguin.
World Bank. 1982. *World development report, 1982*. New York: Oxford University Press.

Poverty Alleviation Programmes

Martin Ravallion

Abstract

This article reviews the issues and evidence concerning a class of policies that aim to reduce poverty by providing direct current relief to those in need and/or by compensating for market and governmental failures that help perpetuate poverty. The article focuses on programmes found in developing countries. Poverty proxies or self-targeting mechanisms are typically used and the specific policies discussed include contingent transfers, community-based programmes, social funds and workfare programmes.

Keywords

Affirmative action; Agency costs; Conditional transfers; Deadweight losses; Equity–efficiency trade-off; Factor mobility; Indicator-based targeting; Intrahousehold welfare; Learning-by-doing; Poverty alleviation programmes; Poverty proxies; Poverty traps; Programme evaluation; Redistribution of income; Workfare

JEL Classifications

O1

Rapid poverty reduction is widely seen to call for a combination of policies that on the one hand promote economic growth and on the other help poor people share in, and contribute to, the opportunities of a growing economy. There is wide agreement that the latter set of policies should include universal provision of adequate basic health care and schooling. There is less agreement on the scope for 'poverty alleviation programmes,' typically entailing transfers in cash or kind targeted to poor people. This article provides an overview of such programmes. First their objectives and the factors constraining their performance are discussed. Then the focus turns to the types of programmes found in developing countries.

Objectives and Constraints

The generally agreed objective of this class of policies is to increase the standard of living of those with low levels of living, that is, to reduce 'absolute poverty'. While recognizing that high levels of inequality can impede prospects for reducing poverty, the objective of this class of policies is poverty reduction, not redistribution per se. Trade-offs underlie this objective. Inequality-reducing interventions can come at a cost to efficiency, such as through effects on the work effort or savings of beneficiaries. While these costs can be serious for specific programmes in specific contexts, it should not be presumed that there will necessarily be an equity–efficiency trade-off. In a world of market failures and 'poverty traps' direct interventions against poverty can also promote aggregate efficiency and (hence) growth. (On how there can be too much inequality and risk from the point of view of aggregate output see, *inter alia,* Bénabou 1996; Aghion et al. 1999; and Bardhan et al. 2000. On poverty traps, see, *inter alia,* Dasgupta 1993; Banerjee and Newman 1994; and Hoff 2001. Policy implications are examined in Ravallion 2005a; World Bank 2001, 2006.) For example, credit constraints leave unexploited investment opportunities, notably for the poor (who have little or no collateral). Agency costs are probably also borne more heavily by the poor. (Agency costs arise when an agent, such as a worker or tenant farmer, makes key decisions relevant to a principal – the capitalist or land owner – who faces high supervision costs. Such models can generate efficient redistributions, from principal to agent; see Bowles and Gintis 1996.)

There can be other trade-offs. The programmes that are best for reducing current poverty need not coincide with those that are best for reducing future poverty; examples will be given later. And the policies that are good for reducing chronic poverty (such as promoting the adoption of new farming technologies) may matter little to, or even exacerbate, transient poverty (by exposing poor farmers to greater downside risk).

There are a number of constraints in formulating effective anti-poverty programmes. Governmental budgets figure prominently. Interventions in the name of poor people that require less public spending on other things that matter to their welfare, or are financed in distortionary or inflationary ways that reduce growth, may well increase poverty. The political economy will also constrain the feasible set of anti-poverty policies. What is feasible in practice will of course depend on the specific context.

The scope for these policies is naturally constrained by the information available and administrative capabilities for acting on that information. Problems of information and incentives are at the heart of programme design. Addressing these problems can increase administrative costs, depleting the net resource transfer to the poor. Informational constraints are particularly relevant in the rural and urban informal sectors of developing countries, where policies such as a progressive income tax are seldom feasible (though such policies are themselves second-best responses to information constraints even in rich countries).

Programmes differ in the emphasis given to enhancing the assets of poor people as opposed to raising their current incomes. In principle, poverty-creating inefficiencies due to credit market failures or agency costs can be ameliorated by asset redistributions. However, governments face political-economy constraints on their ability to redistribute wealth. Certain asset-based interventions tend to be more feasible than others. Reducing inequalities of opportunity by improving the schooling and health of children from poor families is often politically easier than reducing inequalities in the ownership of non-human capital or land. And even when asset redistribution is feasible, state-contingent income transfers may also be needed to help address failures in the provision of private insurance. It is likely that antipoverty policy will continue to call for a mix of efforts to redress inequalities of opportunity (probably emphasizing human resource development) and specific transfers in cash or kind.

Another issue is how finely targeted antipoverty programmes should be. Policy discussions often call for better 'targeting' – a higher share of total spending going to the poor. However, the most finely targeted programme need not be the one with the greatest impact on poverty. Fine targeting can increase administrative costs, yield deadweight losses (as will be illustrated later) and undermine political support for the programme. (On the political economy of targeting see Gelbach and Pritchett 2000. On deadweight losses see Ravallion and Datt 1995. On administrative costs see Grosh 1995. More general discussions of these issues can be found in Besley and Kanbur 1993; van de Walle 1998.) Uncertainties about the measures used in practice to identify the poor can also lead one to question the benefits of fine targeting.

Reliable monitoring and *ex post* evaluation is crucial. Our knowledge about the performance of these programmes has traditionally been poor, but this is changing as more resources and better data and methods go into impact evaluations. (An impact evaluation measures impacts on outcomes relative to explicit counterfactuals.

Ravallion 2005b, reviews methods and results on the impact evaluation of this class of policies.) These have revealed both successes and failures, often depending crucially on the context; the same type of programme can achieve very different outcomes in different settings including at different scales of operation. (Theory and evidence indicating that targeting performance tends to improve as a programme expands can be found in Ravallion 2005c.) Thus greater emphasis is now given to adapting programmes to their context – 'learning-by-doing' – as well as to broader reforms in governance and new, more pro-poor, institutions that can help assure better policy-making and implementation.

The following discussion briefly examines the main types of programmes found in practice, which will illustrate some of the generic points above. While 'targeting' per se is not the objective, existing programmes can usefully be classified according to the way they try to target the poor. The focus is on programmes that rely on transfers in cash or kind. (Lipton and Ravallion 1995, review the full range of antipoverty policies found in practice, which, in addition to transfers, include various forms of direct support to smallholders, better instruments for credit and insurance, tenancy reforms and titling programmes to enhance security of access to land, and removing biases against the poor in taxation, spending and regulatory – including migration – policies.)

Indicator Targeting

The problems of observing incomes and the incentive effects of means-testing have led to various schemes that make transfers in cash or kind according to 'poverty proxies' such as living in a poor area, age (both children and the elderly) and rural landlessness. Everyone with the same value of the indicator (or some combination) is treated the same way. Tools exist for finding optimal allocations to minimize a poverty index based on poverty proxies and for measuring the impact on poverty (Ravallion 1993). Naturally, the more information that is available, the better indicator targeting works. Significant advances have been made in our ability to exploit sample survey information for the purposes of informing policy-making. For example, reasonably reliable and quite detailed 'poverty maps' can now be formed by combining sample survey data with census data; see, for example, Elbers et al. (2003).

Policy-makers have often been overly optimistic about how well they can reach the poor based on readily observable indicators. Here there are some sobering lessons from empirical research. Even using comprehensive, high-quality household sample surveys we have rather modest ability to account for differences in the levels of measured consumption or incomes in terms of the sorts of readily observed covariates that are typically used for targeting. There appears to be sizable heterogeneity in living standards within target groups identified by poverty proxies. Further sources of targeting errors arise from the fact that one must base actual policies on data for the whole population (not just a sample survey) and that respondents will naturally face incentives to distort the data when it is known why it is being collected. Thus, one can expect (possibly large) errors in practice when using indicator targeting to fight poverty.

Performance in reaching the transiently poor through indicator-based targeting appears to be generally worse than performance in reaching the chronically poor (see Ravallion et al. 1995; Lokshin and Ravallion 2000; van de Walle 2004). This is not too surprising given that widely used poverty proxies have even less ability to explain changes over time in levels of living (with the use of panel data in which the same households are interviewed over time). And stakeholders naturally resist changes to a programme's allocations. Despite the potential in theory, targeted transfers in practice have not responded rapidly to changing household circumstances, as would be required for effective insurance.

Such observations have prompted efforts to find 'smart policies' that rely more on incentives in their design and can adapt more rapidly, and on institutional changes that can help assure that poor people are better represented in decision-making; examples are given below.

There are also reasons for thinking that the benefits of indicator targeting are sometimes underestimated. Policy discussions have typically viewed targeting in a static non-behavioural way; location or ethnicity are simply poverty proxies. Recent research has offered a new perspective, pointing to the potential for efficiency gains from targeting groups being locked out of economic opportunities by market or political mechanisms. For example, residential stratification in the presence of externalities can generate persistent inequality (Bénabou 1993; Durlauf 1996). There is evidence of 'geographic poverty traps' in underdeveloped rural economies, such that living in a poorly endowed area reduces prospects of escaping poverty at given individual (non-geographic) characteristics; see Jalan and Ravallion (2002), who use data for China. Poor-area development programmes in such a setting can thus secure long-term gains (Jalan and Ravallion 1998). There is also evidence that the political economy can generate biases against specific groups defined by location or ethnicity and that affirmative actions favouring these groups can enhance the impact of an anti-poverty programme (Besley et al. 2004). Specific demographic groups (both children and the elderly) have also been targeted, given the evidence of strong demographic correlates of poverty found in household survey data, though the robustness of these empirical findings to measurement assumptions is questionable. (Allowing for scale economies in consumption can readily reverse the common finding that larger households tend to be poorer based on consumption or income per person; Lanjouw and Ravallion 1995.) Here too there can be efficiency gains. South Africa has a pension scheme that gives cash transfers to the elderly; Duflo (2003) reports that these pensions have positive external benefits for child health within recipient families. The upshot of these findings is that targeting certain groups can have a greater long-term impact on poverty than suggested by a purely statistical poverty profile.

Finding that transfers based on indicators of current poverty can bring long-term benefits (given factor market imperfections) does not, however, mean that they are the best policy option for this purpose. Policies to increase factor mobility can also have a role. Incentives to attract private capital into poorly endowed areas or to encourage labour migration out of them could well be more poverty reducing than transfers targeted to those areas. There has been very little work on these policy choices, and one often hears unsubstantiated claims by advocates.

Securing the efficiency gains from targeted transfers will often require complementary programmes or reforms. This has been emphasized in the context of redistributive land reforms, where impediments in access to credit and technologies can greatly attenuate the efficiency gains (Binswanger et al. 1995; World Bank 2003). Recognition of the need to combine transfers (of specific assets or incomes) with other initiatives to help foster the productivity of the poor has prompted recent interest in a class of conditional transfers that we now turn to.

Conditional Transfers

Many anti-poverty programmes impose conditions on recipients that attempt to change their behaviour. The (more or less explicit) rationale is that some form of market failure has entailed that current behaviours are not socially optimal. (On the efficiency arguments for conditionality requirements on transfer schemes see Das et al. 2004.) In the 1990s, a number of new conditional transfer programmes emerged that required recipients to satisfy schooling (and sometimes child health-care) requirements. An example is Bangladesh's Food-for-Education (FFE) programme, which relies on community-based targeting of food transfers that aim to create an incentive for reducing the costs to the poor of market failures. Other examples are PROGRESA (renamed Oportunidades) in Mexico and Bolsa Escola in Brazil; in these programmes cash transfers are targeted to certain demographic groups in poor areas, conditional on regular school attendance and visits to health centres.

If one was concerned solely with current income gains to participants then one would not want to make transfers conditional on school attendance, which imposes a cost on poor families by inducing them to withdraw children from the labour force, thus reducing the (net) income gain to the poor from the programme. Rather, this type of programme is aiming to balance a current poverty reduction objective against an objective of reducing future poverty. Given the credit market failure, the incentive effect on labour supply of the programme (often seen as an adverse outcome of transfers) is now judged to be a benefit – to the extent that a well-targeted transfer allows poor families to keep the kids in school rather than sending them to work. Notice too that concerns about distribution *within* the household often underlie the motivation for such programmes; the programme conditionality makes it likely that relatively more of the gains accrue to children. This can also be interpreted as a policy response to the deficiency of traditional poverty proxies in reflecting distribution within the household.

There is evidence of significant gains from Bangladesh's FFE programme in terms of school attendance, with only modest income forgone through displaced child labour (Ravallion and Wodon 2000). The programme was able to appreciably increase schooling, at modest cost to the current incomes of poor families. Mexico's PROGRESA programme has also been found to increase schooling, though the gains appear to be lower than for FFE (Behrman et al. 2002; Schultz 2004; Skoufias 2005). This is probably because primary schooling rates are higher in Mexico, implying less value-added over the (counterfactual) schooling levels that would obtain otherwise. Sadoulet and de Janvry (2002) argue that there would be greater efficiency gains (through higher schooling) from PROGRESA if the programme had concentrated on children less likely to attend school in the absence of the programme, notably by focusing on the transition to secondary school. However, the policy choice will depend critically on the weight one attaches to current income gains for the poor as against future gains through schooling.

Community-Based Programmes

In recent times, community participation in programme design and implementation has been advocated as an institutional change that can help relieve informational constraints, and possibly tilt the balance of power toward the poor. A common form of this idea in practice is that the central government sets up a 'social fund' that provides financial support to a potentially wide range of community-based projects, with strong emphasis on local participation in proposing and implementing the specific projects. Community (governmental or non-governmental) organizations are assumed to be better informed about what is needed. The centre retains control over how much goes to each locality. A useful overview of what is known about this class of programs can be found in Mansuri and Rao (2004).

While 'empowerment' of the poor has motivated such community-based efforts, capture by local elites has been a continuing concern. Reliable generalizations are as yet elusive. There are reasons to expect heterogeneity across communities in the impacts of the same programme. Relevant sources of heterogeneity identified in the literature include local asset inequality (Bardhan and Mookherjee 2000; Galasso and Ravallion 2005) and the extent of interlinkage in local social networks (Spagnolo 1999).

In the design of Bangladesh's FFE programme, economically backward areas were supposed to be chosen by the centre, leaving community groups – exploiting idiosyncratic local information – to select participants within those areas. Galasso and Ravallion (2005) use survey data to assess FFE incidence within and between villages. They found that targeting performance – measured by the difference between the realized per capita allocation to the poor and the non-poor – varied greatly between villages. Higher allocations from the centre to a village tended to yield better targeting performance, but there was no sign that poorer villages were any better or worse at targeting their poor.

The results also point to the role played by antecedent inequalities within villages in determining the relative power of the poor in local decision-making. Galasso and Ravallion found that more unequal villages are worse at targeting the poor – consistent with the view that greater land inequality comes with less power for the poor in village decision-making. (This echoes the view that inequalities can persist through their influence on the institutions that develop. For example, Engerman and Sokoloff 2005, argue that this is why high initial inequality persisted in colonized countries. Also see World Bank 2006.) This suggests a mechanism whereby inequality is perpetuated through the local political economy; the more unequal the initial distribution of assets, the better placed the non-poor will be to capture the benefits of external efforts to help the poor.

Self-Targeting

The informational constraints on anti-poverty programmes have strengthened arguments for using self-targeting mechanisms. There are numerous ways to use incentives in programme design to assure self-targeting of the poor. For example, the rationing of food or health subsidies by queuing can be self-targeting (Alderman 1987), as can subsidizing inferior food staples or packaging in ways that are unappealing to the non-poor. However, the classic example of a self-targeted anti-poverty programme is a workfare programme, in which work requirements are imposed on welfare recipients with the aim of creating incentives to encourage participation only by the poor and to reduce dependency on the programme. (Besley and Coate 1992, provide a formal model of the incentive arguments.)

An example is the famous Employment Guarantee Scheme (EGS) in Maharashtra, India. This aims to assure income support in rural areas by providing unskilled manual labour at low wages to anyone who wants it. The scheme is financed domestically, largely from taxes on the relatively well-off segments of Maharashtra's urban populations. The employment guarantee helps support the insurance function, and is also seen to help empower poor people. In practice, however, most workfare schemes have entailed some administrative rationing of the available work, often in combination with geographic targeting.

Workfare schemes generally have a good record in screening the poor from the non-poor, and providing effective insurance against both covariate and idiosyncratic shocks. (For evidence on this point see Ravallion and Datt 1995; Subbarao 1997; Jalan and Ravallion 2003; Coady et al. 2004.) They have provided protection when there is a threat of famine (Drèze and Sen 1989) or in the wake of a macroeconomic crisis (see, for example, Pritchett et al. 2003, for Indonesia's crisis in 1998, and Galasso and Ravallion 2004, for Argentina's crisis in 2002). Design features are crucial, notably that the wage rate is not set too high. For example, Ravallion et al. (1993) provide evidence on how the EGS responds to aggregate shocks, and on how its ability to insure the poor was jeopardized by a sharp increase in the wage rate. Low-wage workfare schemes have been advocated as a core element of a 'permanent safety net' for risk-prone economies (Ravallion 2005c).

Self-targeted schemes face a trade-off between targeting performance (meaning their ability to concentrate benefits on the poor) and net income gains to participants, given that these programmes work by deliberately imposing costs on participants. Self-targeting requires that the cost of participation is higher for the non-poor than the poor (so that it is the poor who tend to participate), but it may not be inconsequential for the poor.

A potentially important cost to workfare participants is forgone income. This is unlikely to be zero; the poor can rarely afford to be idle. An estimate for two villages in Maharashtra, India, found that the forgone income from employment on the EGS was quite low – around one quarter of gross wage earnings; most of the time displaced was in domestic labour, leisure and unemployment (Datt and Ravallion 1994). By contrast, for Argentina's Trabajar Program (a combination of workfare and social fund), it was estimated that about one half of gross wage earnings was taken up by forgone incomes (Jalan and Ravallion 2003; Ravallion et al. 2005).

Workfare schemes also illustrate the potential trade-off in policy design between short-term income gains to the poor and longer-term gains through asset creation. Workfare programmes have not traditionally emphasized the value to the poor of the assets created, which appear often to mainly benefit the non-poor or to be of remarkably little value to anyone (see, for example, Gaiha 1996, writing about Maharashtra's EGS.) The Trabajar Program illustrates the scope for a new wave of workfare programmes that emphasize asset creation in poor communities.

The programme's design gave explicit incentives (through the *ex ante* project selection process) for targeting the asset creation to poor areas, again compensating for the market failures that help create poor areas in the first place. There is typically much useful work to do in poor neighbourhoods – work that would probably not get financed otherwise.

The choice between the goal of raising current incomes of the poor and reducing future poverty will never be a straightforward. The choice will naturally depend on circumstances. For example, in macroeconomic or agro-climatic crises it is to be expected that the emphasis will shift to current income gains, away from asset creation – implying, for example, more labour-intensive sub-projects on workfare programmes.

See Also

- ▶ Income Taxation and Optimal Policies
- ▶ Poverty Lines
- ▶ Social Insurance

Bibliography

Aghion, P., E. Caroli, and C. Garcia-Penalosa. 1999. Inequality and economic growth: The perspectives of the new growth theories. *Journal of Economic Literature* 37: 1615–1660.

Alderman, H. 1987. Allocation of goods through non-price mechanisms: Evidence on distribution by willingness to wait. *Journal of Development Economics* 25: 105–124.

Banerjee, A., and A. Newman. 1994. Poverty, incentives and development. *American Economic Review: Papers and Proceedings* 84: 211–215.

Bardhan, P., and D. Mookherjee. 2000. Capture and governance at local and national levels. *American Economic Review: Papers and Proceedings* 90: 135–139.

Bardhan, P., S. Bowles, and H. Gintis. 2000. Wealth inequality, wealth constraints and economic performance. In *Handbook of income distribution*, vol. 1, ed. A. Atkinson and F. Bourguignon. Amsterdam: North-Holland.

Behrman, J., Sengupta, P. and Todd, P. 2002. Progressing through PROGRESSA: An impact assessment of a school subsidy program. Mimeo: University of Pennsylvania.

Bénabou, R. 1993. Workings of a city: Location, education and production. *Quarterly Journal of Economics* 108: 619–652.

Bénabou, R. 1996. Inequality and growth. In *National Bureau of Economic Research macroeconomics annual*, ed. B. Bernanke and J. Rotemberg. Cambridge, MA: MIT Press.

Besley, T., and S. Coate. 1992. Workfare vs. welfare: Incentive arguments for work requirements in poverty alleviation programs. *American Economic Review* 82: 249–261.

Besley, T., and R. Kanbur. 1993. Principles of targeting. In *Including the poor*, ed. M. Lipton and J. van der Gaag. Washington, DC: World Bank.

Besley, T., R. Pande, L. Rahman, and V. Rao. 2004. The politics of public good provision: Evidence from Indian local governments. *Journal of the European Economic Association* 2: 416–426.

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

Bowles, S., and H. Gintis. 1996. Efficient redistribution: New rules for markets, states and communities. *Politics and Society* 24: 307–342.

Coady, D., M. Grosh, and J. Hoddinott. 2004. Targeting outcomes redux. *World Bank Research Observer* 19: 61–86.

Das, J., Q.-T. Do, and B. Ozler. 2004. A welfare analysis of conditional cash transfer schemes: Implications for policy. *World Bank Research Observe* 20: 57–80.

Dasgupta, P. 1993. *An inquiry into well-being and destitution*. Oxford: Oxford University Press.

Datt, G., and M. Ravallion. 1994. Transfer benefits from public works employment: Evidence from rural India. *Economic Journal* 104: 1346–1369.

Drèze, J., and A. Sen. 1989. *Hunger and public action*. Oxford: Oxford University Press.

Duflo, E. 2003. Grandmothers and granddaughters: Old age pension and intrahousehold allocation in South Africa. *World Bank Economic Review* 17: 1–26.

Durlauf, S. 1996. A theory of persistent income inequality. *Journal of Economic Growth* 1: 75–93.

Elbers, C., P. Lanjouw, and J. Lanjouw. 2003. Micro-level estimation of poverty and inequality. *Econometrica* 71: 355–364.

Engerman, S. and K. Sokoloff. 2005. Colonialism, inequality and long-run paths of development. Working Paper No. 11057. Cambridge, MA: NBER.

Gaiha, R. 1996. How dependent are the rural poor on employment guarantee scheme in India? *Journal of Development Studies* 32: 669–694.

Galasso, E., and M. Ravallion. 2004. Social protection in a crisis: Argentina's Plan Jefes y Jefas. *World Bank Economic Review* 18: 367–399.

Galasso, E., and M. Ravallion. 2005. Decentralized targeting of an anti-poverty program. *Journal of Public Economics* 85: 705–727.

Gelbach, J., and L. Pritchett. 2000. Indicator targeting in a political economy: Leakier can be better. *Journal of Policy Reform* 4: 113–145.

Grosh, M. 1995. Toward quantifying the trade-off: Administrative costs and incidence in targeted programs in

Latin America. In *Public spending and the poor: Theory and evidence*, ed. D. van de Walle and K. Nead. Baltimore: Johns Hopkins University Press.

Hoff, K. 2001. Beyond Rosenstein-Rodan: The modern theory of coordination problems in development. In *Proceedings of the annual world bank conference on development economics 2000*. Washington, DC: World Bank.

Jalan, J., and M. Ravallion. 1998. Are there dynamic gains from a poor-area development program? *Journal of Public Economics* 67: 65–86.

Jalan, J., and M. Ravallion. 2002. Geographic poverty traps? A micro model of consumption growth in rural China? *Journal of Applied Econometrics* 17: 329–346.

Jalan, J., and M. Ravallion. 2003. Estimating the benefit incidence of an anti-poverty program by propensity-score matching. *Journal of Business and Economic Statistics* 21: 19–30.

Lanjouw, P., and M. Ravallion. 1995. Poverty and household size. *Economic Journal* 105: 1415–1435.

Lipton, M., and M. Ravallion. 1995. Poverty and policy. In *Handbook of development economics*, vol. 3, ed. J. Behrman and T. Srinivasan. Amsterdam: North-Holland.

Lokshin, M., and M. Ravallion. 2000. Welfare impacts of Russia's 1998 financial crisis and the response of the public safety net. *Economics of Transition* 8: 269–295.

Mansuri, G., and V. Rao. 2004. Community-based and -driven development. *World Bank Research Observer* 19: 1–40.

Pritchett, L., S. Sumarto, and A. Suryahadi. 2003. Safety nets or safety ropes? Dynamic benefit incidence of two crisis programs in Indonesia. *World Development* 31: 1257–1277.

Ravallion, M. 1993. Poverty alleviation through regional targeting: A case study for Indonesia. In *The economics of rural organization: Theory, practice and policy*, ed. K. Hoff, A. Braverman, and J. Stiglitz. Oxford: Oxford University Press.

Ravallion, M. 2005a. Transfers and safety nets in poor countries: Revisiting the tradeoffs and policy options. In *Understanding poverty*, ed. A. Banerjee, R. Bénabou, and D. Mookerjee. New York: Oxford University Press.

Ravallion, M. 2005b. Evaluating anti-poverty programs. In *Handbook of agricultural economics*, vol. 4, ed. R. Evenson and T. Schultz. Amsterdam: North-Holland.

Ravallion, M. 2005c. Who is protected? On the incidence of fiscal adjustment. In *Macroeconomic policies and poverty reduction*, ed. A. Mody and C. Pattillo. London: Routledge.

Ravallion, M., and G. Datt. 1995. Is targeting through a work requirement efficient? In *Public spending and the poor: Theory and evidence*, ed. D. van de Walle and K. Nead. Baltimore: Johns Hopkins Press.

Ravallion, M., and Q. Wodon. 2000. Does child labor displace schooling? Evidence on behavioral responses to an enrolment subsidy. *Economic Journal* 110: C158–C176.

Ravallion, M., G. Datt, and S. Chaudhuri. 1993. Does Maharashtra's employment guarantee scheme guarantee employment? Effects of the 1988 wage increase. *EconomicDevelopment and Cultural Change* 41: 251–275.

Ravallion, M., D. van de Walle, and M. Gautam. 1995. Testing a social safety net. *Journal of Public Economics* 57: 175–199.

Ravallion, M., E. Galasso, T. Lazo, and E. Philipp. 2005. Do workfare participants recover quickly from retrenchment? *Journal of Human Resources* 40: 208–230.

Sadoulet, E. and A. de Janvry. 2002. Alternative targeting and calibration schemes for educational grants programs: Lessons from PROGRESA. Mimeo. Berkeley: University of California.

Schultz, T. 2004. School subsidies for the poor: Evaluating the Mexican PROGRESA poverty program. *Journal of Development Economics* 74: 199–250.

Skoufias, E. 2005. PROGRESA and its impact on the welfare of rural households in Mexico. Research Report No. 139. Washington, DC: International Food Research Institute.

Spagnolo, G. 1999. Social relations and cooperation in organizations. *Journal of Economic Behavior and Organization* 38: 1–25.

Subbarao, K. 1997. Public works as an anti-poverty program: An overview of crosscountry experience. *American Journal of Agricultural Economics* 79: 678–683.

van de Walle, D. 1998. Targeting revisited. *World Bank Research Observer* 13: 231–248.

van de Walle, D. 2004. Testing Vietnam's public safety net. *Journal of Comparative Economics* 32: 661–679.

World Bank. 2001. *World development report: Attacking poverty*. New York: Oxford University Press.

World Bank. 2003. *Land policies for growth and poverty reduction*. New York: Oxford University Press.

World Bank. 2004. *World development report: Delivering public services to the poor*. New York: Oxford University Press.

World Bank. 2006. *World development report: Equity and development*. New York: Oxford University Press.

Poverty Lines

Martin Ravallion

Abstract

The article provides welfare-economic definitions of poverty lines and critically assesses the main methods of setting poverty lines found in practice. These can be interpreted as ways of expanding the information set used in applied work to address some long-standing problems in measuring welfare. Objective methods draw on information from outside economics on the commodities needed for normative activity levels. Subjective methods extend the information base by drawing on selfreported perceptions of consumption adequacy, allowing estimation of an endogenous social subjective poverty line.

Keywords

Consumers’ expenditure; Cost-of-basic-needs poverty measure; Equivalent income functions; Food-energy-intake poverty measure; Objective poverty lines; Poverty; Poverty lines; Relative income; Relative poverty line; Rowntree, B.; Self-assessed welfare; Social subjective poverty line; Subjective poverty lines; Utility; Welfare economics; Well-being

JEL Classifications

O1

Knowing how many people live in households with income or consumption expenditure below the ‘poverty line’ has helped focus attention on the extent of poverty and has informed policymaking for fighting poverty. But how are poverty lines defined and calculated? This article first provides a theoretical definition, and then describes the main methods found in practice.

Poverty Lines in Theory

People in different circumstances – with different household sizes or demographic compositions or living in different places – naturally have different levels of economic welfare at the same level of income. They have different needs. A poverty line should reflect these differences. But how should that be done? To answer this question we must first define the conceptual ideal against which the methods found in practice are to be judged.

The poverty line for a given individual can be defined as the money the individual needs to achieve the minimum level of ‘welfare’ to not be deemed ‘poor’, given its circumstances. Everyone at the poverty line is taken to be equally badly off, and all those below the line are worse off than all above it. The next question is: what concept of ‘welfare’ should serve as the anchor for the poverty line? For economists the obvious answer is ‘utility’. A justification for utility-consistent poverty lines can be found by applying standard welfare-economic principles to poverty measurement (Ravallion 1994). These principles are that assessments of social welfare should depend solely on utilities, that people with the same initial utility should be treated the same way, and that social welfare should not be decreasing in any utility.

To formalize this definition, consider individual i with characteristics x_i (a vector). The interpersonally comparable utility function is $u(q_i,x_i)$. The quantity vector q_i is utility maximizing, giving demand functions $q(p_i,y_i, x_i)$ at total expenditure y_i and corresponding utility maximum, $v(p_i,y_i, x_i)$. The utility-consistent poverty line is the point on the consumer’s expenditure function corresponding to a common reference utility level. (As always, the monetary measurement of welfare requires a reference utility level.) The consumer’s expenditure function is $e(p_i, x_i, u)$, giving the minimum cost of utility u when facing the price vector p_i. Let u_z denote the minimum utility deemed necessary to escape poverty. Consistency requires that this is a constant across all i. The money metric of u_z defines the *utility-consistent poverty lines:*

$$z_i^u = e(p_i, x_i, u_z) \text{ for all } i = 1, \dots n \qquad (1)$$

This is closely related to a number of other economic concepts. Eq. (1) is ‘money-metric utility’ at a specific reference utility, interpretable as the poverty line in utility space. (Textbook treatments of money-metric utility functions – sometimes called ‘equivalent income functions’ – can be found in Varian 1978, and Deaton and Muellbauer 1980.) The value of $z_i^u/e(p_r, x_r, u_z)$ for reference individual r gives the ‘true cost-of-living index’ (see, for example, Deaton and Muellbauer 1980). The value of y_i/z_i^u is the 'welfare ratio' (Blackorby and Donaldson, 1987). On exploiting the properties of the expenditure function, Eq. (1) can be written in a more instructive form:

$$z_i^u = p_i q(p_i, x_i, u_z) \tag{2}$$

Thus, the poverty line is the cost of a bundle of goods, namely, the vector of utility-compensated (Hicksian) demands, $q(p_i, x_i, u_z)$. This bundle can be interpreted as 'basic consumption needs'.

Note that an absolute poverty line in terms of utility can have the properties of a *relative poverty line* in the income space, in that it rises with the mean income of a relevant reference group. This is possible if individual utility depends on both 'own income' and income relative to others in that reference group. In other words, the indirect utility function has the form, $v\left(p_i, y_i, y_i/\overline{y}_i^r, x_i\right)$ where $\overline{y}_i^r$ is the mean income of the reference group(s); the poverty line takes the form $z_i = z\left(p_i, \overline{y}_i^r, x_i, u_z\right)$ (which solves $u_z = v\left(p_i, z_i, z_i/\overline{y}_i^r, x_i\right)$). Thus, we can view poverty as absolute in the space of welfare, but relative in the space of commodities (paraphrasing Sen 1983). However, the extreme case in which the poverty line is a constant proportion of the mean (as used by poverty measures derived by Eurostat, for the European Commission) requires the seemingly implausible assumption that utility depends *solely* on relative income, given by the ratio of own income to the mean. By this measure, if all incomes increase by the same multiplicative factor then the proportion of people living below the poverty line will be unchanged. Cross-country comparisons of poverty lines suggest that relative income is valued more highly as mean income rises; in the poorest countries, absolute income levels are the dominant consideration, so then the poverty lines tend to have a low elasticity to the mean (Ravallion 1994, 1998). There is micro empirical evidence supporting this view, based on self-assessed welfare (Ravallion and Lokshin 2005).

For economists, utility is the obvious anchor for setting poverty lines, but it is not the only approach. Functioning-based concepts of welfare offer an alternative foundation for poverty lines. The approach can be characterized in the terms of Sen's (1985) argument that 'well-being' should be thought of in terms of a person's capabilities, that is, the functionings ('beings and doings') that a person is able to achieve. On this view, poverty means not having an income sufficient to support specific normative functionings. Utility – as the attainment of personal satisfaction – can be viewed as one such functioning relevant to well-being (Sen 1992, ch. 3). But it is possibly only one of the functionings that matter. Independently of utility, one might say that a person is better off if she is able to participate fully in social and economic activity.

From this starting point, a more general theoretical formalization of the definition of a poverty line can be proposed as follows. Let a person's functionings be determined by the goods she consumes and her characteristics, giving the vector of functionings:

$$f_i = f(q_i, x_i) \tag{3}$$

where f is a vector-valued function. One can postulate that a person derives utility directly from her functionings. We can then interpret $u(q_i, x_i)$ as a derived utility function, obtained by substituting (3) into the (primal) utility function defined over functionings.

Functioning consistency for a set of poverty lines requires that certain normative functionings are reached at the poverty line. Let f_z denote the vector of critical functionings needed to not be deemed poor. These are normative judgements, just as u_z is a normative judgment. Assume that there is a bundle q_i^c such that no functioning is below its critical value:

$$f_z \leq f\left(q_i^c, x_i\right) \tag{4}$$

This yields the poverty lines: $z_i^c = p_i q_i^c$. There can be multiple solutions for q_i^c. Two ways to pick a unique poverty line can be identified. The first is to define z_i^c as the minimum y such that (4) holds. Notice that one (or more) specific functionings will be decisive; that is, the functioning that is the last to reach its critical value as income rises. In this sense, the lowest priority functioning for the individual will be decisive. The alternative approach is to treat attainments as a random variable (that is, with a probability distribution) and take a mean conditional on income and other identified covariates, including group membership. Then poverty lines are deemed to be functioning consistent if f_z is reached in expectation.

Implementing these concepts empirically requires that we solve two problems. The first can be called the *referencing problem:* what is the reference level of utility that anchors the poverty line (u_z in Eq. 1)? It is tempting to say this choice is arbitrary, and to hope that it is innocuous. But the choice of the reference is far from arbitrary, and (in general) it affects the resulting poverty measure. This speaks to the importance of testing the sensitivity of poverty comparisons (such as between groups or over time) to the choice of reference, as it determines the level of the poverty line. Tests exist for the robustness of ordinal comparisons using stochastic dominance criteria; on this approach see Atkinson (1987) and Ravallion (1994).

The second problem is the *identification problem.* Even if we can readily agree on what the poverty line is in welfare space, there is a further problem in identifying the expenditure function in Eq. (1). Standard practice is to calibrate the parameters of the cost function from consumer demand behaviour. The problem is that individuals vary in characteristics, such as their size and demographic composition, which influence welfare in ways that may not be evident in consumer demand behaviour. Then there is a fundamental problem of identification (Pollak and Wales 1979; Pollak 1991).

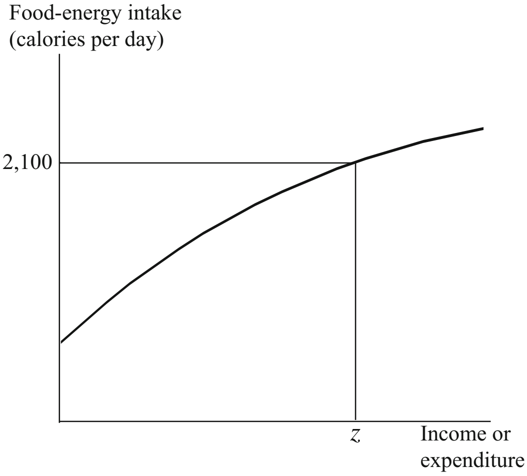
Poverty Lines in Practice

The methods of setting poverty lines found in practice fall under two headings: objective poverty lines and subjective poverty lines.

Objective Poverty Lines

The main methods found in practice are the food-energy-intake (FEI) method and the cost-of-basic-needs (CBN) method. It is known that these methods give radically different results; using data for Indonesia, Ravallion and Bidani (1994) found virtually zero correlation between the regional poverty profiles (given the poverty rates across geographic areas) produced by these two methods. Since policy choices (such as in regional targeting) could depend critically on which method is used, it is important to probe carefully into the choice.

The FEI method can be interpreted as a special case of the functioning-based approach described above. The specialization is to focus on just one functioning, namely food-energy intake. The method finds the consumption expenditure or income level at which food-energy intake is just sufficient to meet predetermined food-energy requirements for good health and normal activity levels. (Such caloric requirements are given in WHO 1985, for example.) To deal with the fact that food-energy intakes naturally vary at a given income level, the FEI method typically calculates an expected value of intake at given income. Figure 1 illustrates the method. The vertical axis is food-energy intake, plotted against income (or expenditure) on the horizontal axis. A line of ‘best fit’ is indicated; this is the expected value of caloric intake at given income (that is, the non-linear regression function). By simply inverting this line, one finds the income z at which a person typically attains the stipulated food-energy requirement. This method, or something similar,



Poverty Lines, Fig. 1 The food-energy intake method of setting poverty lines has been used often, including by Dandekar and Rath (1971), Osmani (1982), Greer and Thorbecke (1986), and Paul (1989), and by numerous governmental statistics offices. It is often found in practice in developing countries.

One concern about this method is that the resulting poverty lines need not be consistent in terms of utility or capabilities more generally (Ravallion 1994; Ravallion and Lokshin 2006). Consider first how FEI poverty lines respond to differences in relative prices, which can of course differ across the subgroups (such as regions) being compared in the poverty profile and over time. For example, the prices of many non-food goods relative to food are likely to be lower in urban than in rural areas. This will probably mean that the demand for food and (hence) food-energy intake will be lower in urban than in rural areas, at any given real income.But this does not, of course, mean that urban households are poorer. The relationship between food-energy intake and income will shift according to differences in tastes, activity levels and publicly provided goods. There is nothing in the FEI method to guarantee that these differences are ones that would normally be considered relevant to assessing welfare. Indeed, it is quite possible to find that the 'richer' sector (by the agreed metric of utility) tends to spend so much more on each calorie that it is deemed to be the 'poorer' sector. That has been found to be the case in studies of the properties of FEI poverty profiles for Indonesia (Ravallion and Bidani 1994) and Bangladesh (Ravallion and Sen 1996; Wodon 1997).

Problems also arise in comparisons over time. Suppose that all prices increase, so the cost of a given utility must rise. There is nothing to guarantee that the FEI-based poverty line will increase. That will depend on how relative prices and tastes change; the price changes may well encourage people to consume cheaper calories, and so the FEI poverty line will fall. Wodon (1997) gives an example of this problem in data for Bangladesh. The FEI poverty line fell over time even though prices generally increased. The potential utility inconsistencies in FEI poverty lines are worrying when there is mobility across the subgroups of the poverty profile, such as due to inter-regional migration. For example, it is possible that a process of economic development through urban sector enlargement, in which none of the poor are any worse off and at least some are better off, would result in a measured *increase* in poverty.

The CBN method stipulates a consumption bundle deemed to be adequate for 'basic consumption needs', and then estimates its cost for each of the subgroups being compared in the poverty profile. This is the approach of Rowntree (1901) in his seminal study of poverty in York, England, in 1899, and there have been numerous examples since, including the official poverty lines for the United States (Orshansky 1963; also see Citro and Michael 1995). Some form of functioning consistency is assured by construction, since various valued functionings are essentially the starting point for defining 'basic consumption needs'. The poverty bundle is typically anchored to food-energy requirements consistent with common diets in the specific context. However, allowances for non-food goods are also included, to assure that basic non-nutritional functionings are assured.

The CBN method is utility consistent *if* the right bundle is used, corresponding to the relevant points on the utility-compensated demand functions (Eq. 2). However, there is nothing to guarantee that the bundles of goods built into CBN poverty lines lie on the compensated demand functions, at the (common) reference level of utility. Thus it is important to have some way of assessing a set of CBN poverty bundles. Ravallion and Lokshin (2006) propose an approach to testing the utility consistency of CBN poverty lines across households with common preferences using Samuelson's (1938) theory of revealed preference. However, this can be applied only within subgroups deemed to have common preferences. In practice utility functions can vary, due to differences in climate, for example.

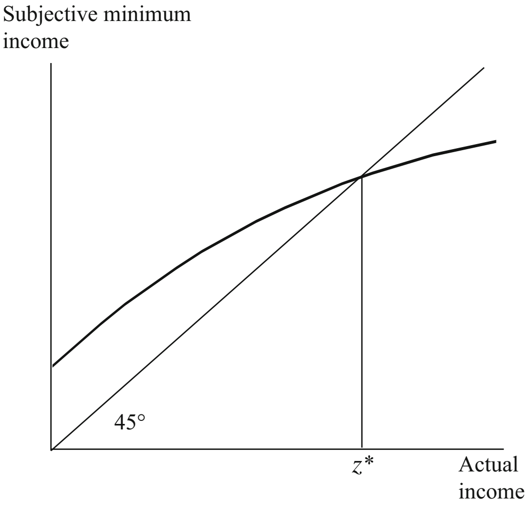
In some cases a complete vector of normative (food and non-food) goods is set, as in Russia's poverty lines (Ravallion and Loskhin 2006). However, it is more often the case that only food needs are set, based on nutritional requirements. To include an allowance for non-food needs, a common practice is to divide the food poverty line by some budget share for food. For example, the US poverty line assumes a food share of one third, so the total poverty line is three times the food line (Orshansky 1963). However, the basis for setting a food share is rarely transparent. Why use the average share, as in the US line? Whose food share should be used?

Arguably, a more appealing approach is to set an allowance for non-food goods that is consistent with demand behaviour at (or in a region of) the food poverty line. Ravallion (1994) proposes two methods. The first divides the food component of the poverty line by the mean food share of households whose actual food spending is in a neighbourhood of the food poverty line. The second method uses mean non-food spending of households whose total sending is in a neighbourhood of the food poverty line. Ravallion argues the first method gives a reasonable upper bound to the allowance for non-food needs while the second gives a lower bound.

Subjective Poverty Lines

There is an inherent subjectivity and social specificity to any notion of 'basic needs', including nutritional requirements. Psychologists, sociologists and others have argued that the circumstances of the individual relative to others influence perceptions of well-being at any given level of individual command over commodities. (Runciman 1966, provided an influential exposition, and supportive evidence. Also see the discussions in Easterlin 1995, and Oswald 1997.) By this view, 'the dividing line . . . between necessities and luxuries turns out to be not objective and immutable, but socially determined and ever changing' (Scitovsky 1978, p. 108).

Subjective poverty lines have been based on answers to the 'minimum income question' (MIQ), such as the following (paraphrased from Kapteyn et al. 1988): 'What income level do you personally consider to be absolutely minimal? That is to say that with less you could not make ends meet.' (This can be thought of as a special case of Van Praag's 1968, 'income evaluation question', which asks what income is considered 'very bad', 'bad', 'not good', not bad', 'good', 'very good'.) One might define as poor all whose actual income is less than the amount they give as an answer to this question. However, this would almost certainly lead to inconsistencies in the resulting poverty measures, in that people with the same income, or some other agreed measure of economic welfare, will be treated differently. Clearly an allowance must be made for heterogeneity, such that people at the same standard of living may well give different answers to the MIQ, but must be considered equally 'poor' for consistency. Past empirical work has found that the expected value of the answer to the MIQ conditional on actual income tends to be an increasing function of actual income. (Contributions include Groedhart et al.1977; Danziger et al.1984; and Kapteyn et al.1988.) Furthermore, past studies have tended to find a relationship such as that depicted in Fig. 2, which gives a stylized representation of the regression function on income for answers to the MIQ. The point z^* in the figure is an obvious candidate for a poverty line; people with income above z^* tend to feel that their income is adequate, while those below z^* tend to feel that it is not. We can call z^* the 'social subjective poverty line' (SSPL).



Poverty Lines, Fig. 2 The social subjective poverty line (z^*)

It is recognized in the literature that there are other determinants of economic welfare which should shift the SSPL, such as family size and demographic composition. Indeed, the answers to the MIQ are interpretable as points on the consumer's expenditure function at a point of minimum utility (Eq. 1). Under this interpretation, subjective welfare assessments provide a means of overcoming the well-known problem of identifying utility from demand behavior alone when household attributes vary (Kapteyn 1994).

While the MIQ has been applied in a number of OECD countries, there have been few attempts to apply it in a developing country. There are a number of potential pitfalls. 'Income' is not a well-defined concept in most developing countries, particularly (but not only) in rural areas. It is not at all clear whether one could get sensible answers to the MIQ. The qualitative idea of the 'adequacy' of consumption is a more promising one in a developing-country setting, and (arguably) many developed counties.

Pradhan and Ravallion (2000) propose a method for estimating the SSPL based on qualitative data on consumption adequacy, as given by responses to appropriate survey questions. Instead of asking respondents what the precise minimum consumption is that they need, one simply asks whether their current consumptions are adequate. This provides a multidimensional extension to the one-dimensional MIQ. The SSPL is the level of total spending above which respondents say (on average) that their expenditures are adequate for their needs. For empirical implementation, the probability that a sampled household will respond that its actual consumption of each type of commodity is adequate can be modelled as a probit regression. Under certain technical conditions, a unique solution for the subjective poverty line can then be obtained from the estimated parameters of the probit regressions for consumption adequacy. Pradhan and Ravallion provide empirical examples for Jamaica and Nepal; the SSPL gave a similar overall poverty rate to preexisting objective poverty lines for both countries, though the structure of the poverty profile was different in some respects: for example, while the objective poverty lines implied that larger households tended to be poorer, this was not the case with the subjective approach.

Subjective data also offer a test of objective poverty lines, by regressing selfrated welfare on income normalized by the poverty line *plus* the variables that went into the construction of the poverty line, which should be jointly insignificant if those lines accord with subjective welfare. This approach is outlined in Ravallion and Lokshin (2002) and illustrated using Russia's poverty lines.

See Also

- ▶ Consumer Expenditure
- ▶ Consumer Expenditure (New Developments and the State of Research)
- ▶ Inflation Measurement
- ▶ Poverty
- ▶ Poverty Alleviation Programmes

Bibliography

Atkinson, A. 1987. On the measurement of poverty. *Econometrica* 55: 749–764.

Blackorby, C., and D. Donaldson. 1987. Welfare ratios and distributionally sensitive cost-benefit analysis. *Journal of Public Economics* 34: 265–290.

Browning, M. 1992. Children and household economic behavior. *Journal of Economic Literature* 30: 1434–1475.

Citro, C., and R. Michael. 1995. *Measuring poverty: A new approach*. Washington, DC: National Academy Press.

Dandekar, V., and N. Rath. 1971. *Poverty in India*. Pune: Indian School of Political Economy.

Danziger, S., J. van der Gaag, E. Smolensky, and M. Taussig. 1984. The direct measurement of welfare levels: How much does it take to make ends meet. *Review of Economics and Statistics* 66: 500–505.

Deaton, A., and J. Muellbauer. 1980. *Economics and consumer behavior*. Cambridge: Cambridge University Press.

Easterlin, R. 1995. Will raising the incomes of all increase the happiness of all? *Journal of Economic Behavior and Organization* 27: 35–47.

Greer, J., and E. Thorbecke. 1986. A methodology for measuring food poverty applied to Kenya. *Journal of Development Economics* 24: 59–74.

Groedhart, T., V. Halberstadt, A. Kapteyn, and B. van Praag. 1977. The poverty line: Concept and measurement. *Journal of Human Resources* 12: 503–520.

Kapteyn, A. 1994. The measurement of household cost functions: Revealed preference versus subjective measures. *Journal of Population Economics* 7: 333–350.

Kapteyn, A., P. Kooreman, and R. Willemse. 1988. Some methodological issues in the implementation of subjective poverty definitions. *Journal of Human Resources* 23: 222–242.

Orshansky, M. 1963. Children of the poor. *Social Security Bulletin* 26: 3–29.

Osmani, S. 1982. *Economic inequality and group welfare*. Oxford: Oxford University Press.

Oswald, A. 1997. Happiness and economic performance. *Economic Journal* 107: 1815–1831.

Paul, S. 1989. A model of constructing the poverty line. *Journal of Development Economics* 30: 129–144.

Pollak, R. 1991. Welfare comparisons and situation comparisons. *Journal of Econometrics* 50: 31–48.

Pollak, R., and T. Wales. 1979. Welfare comparison and equivalence scale. *American Economic Review* 69: 216–221.

Pradhan, M., and M. Ravallion. 2000. Measuring poverty using qualitative perceptions of consumption adequacy. *Review of Economics and Statistics* 82: 462–471.

Ravallion, M. 1994. *Poverty comparisons*. Chur: Harwood Academic Press.

Ravallion, M. 1998. Poverty lines in theory and practice. Living Standards Measurement Study Working Paper No. 133. Washington, DC: World Bank.

Ravallion, M., and B. Bidani. 1994. How robust is a poverty profile? *World Bank Economic Review* 8: 75–102.

Ravallion, M., and M. Lokshin. 2002. Self-rated economic welfare in Russia. *European Economic Review* 46: 1453–1473.

Ravallion, M., and M. Lokshin. 2005. *Who cares about relative deprivation? policy research working paper 3782*. Washington, D.C.: World Bank.

Ravallion, M., and M. Lokshin. 2006. Testing poverty lines. *Review of Income and Wealth* 52: 399–421.

Ravallion, M., and B. Sen. 1996. When method matters: Monitoring poverty in Bangladesh. *Economic Development and Cultural Change* 44: 761–792.

Runciman, W. 1966. *Relative deprivation and social justice*. London: Routledge and Kegan Paul.

Rowntree, B. 1901. *Poverty: A study of town life*. London: Macmillan.

Samuelson, P. 1938. A note on the pure theory of consumer behaviour. *Economica* 5: 61–71.

Scitovsky, T. 1978. *The joyless economy.* Oxford: Oxford University Press.

Sen, A. 1983. Poor, relatively speaking. *Oxford Economic Papers* 35: 153–169.

Sen, A. 1985. *Commodities and capabilities*. Amsterdam: North-Holland.

Sen, A. 1992. *Inequality re-examined*. Oxford: Oxford University Press.

Van Praag, B. *1968. Individual welfare functions and consumer behavior.* Amsterdam: North-Holland.

Varian, H. 1978. *Microeconomic analysis*. New York: Norton.

WHO (World Health Organization). 1985. *Energy and protein Requirements*, Technical Report Series 724. Geneva: WHO.

Wodon, Q. 1997. Food energy intake and cost of basic needs: Measuring poverty in Bangladesh. *Journal of Development Studies* 34: 66–101.

Poverty Traps

Kiminori Matsuyama

Abstract

A poverty trap is a self-perpetuating condition, in which an economy suffers from persistent underdevelopment, vicious circle of poverty, created by circular causation due to the presence of some external economies and/or strategic complementarities. We discuss the concept in a dynamic setting, and review some models of poverty traps in the literature. The policy prescriptions of such models should be treated with caution, since each model identifies one cause; but as many causes are likely to coexist, attempts to pull an economy out of one trap may push it into another.

Keywords

Poverty traps; Stochastic shocks; Human capital; division of labour; Market size; Distortions

JEL Classifications

O0

A poverty trap is a self-perpetuating condition whereby an economy, caught in a vicious circle, suffers from persistent underdevelopment. Although it is often modelled as a low-level equilibrium in a static model of coordination failures, we discuss the concept in a dynamic setting. This is because, in a static setting, we would be unable to distinguish poverty traps from (possibly temporary) bad market outcomes, such as recessions and financial crises, that are also often modelled as low-level equilibriums in a static model of coordination failures.

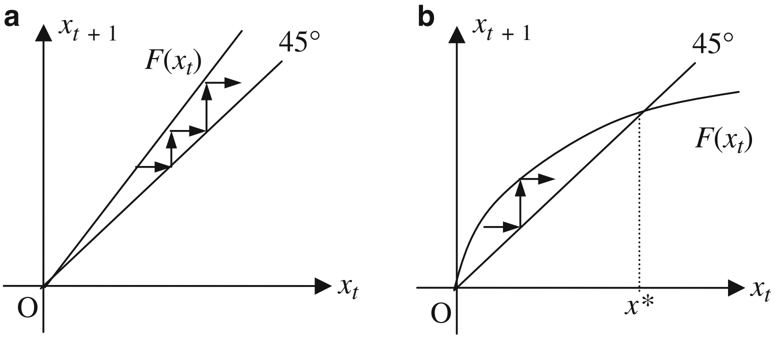
On the Mechanics of Poverty Traps

Imagine that the state of the economy in period t is represented by a single variable, x_t, where a higher x means that the economy is more developed, and that the equilibrium path follows a deterministic one-dimensional difference equation, $x_{t+1} = F(x_t)$. Once the initial condition, x_0, is given, this law of motion can be applied iteratively to obtain the entire trajectory of the economy.

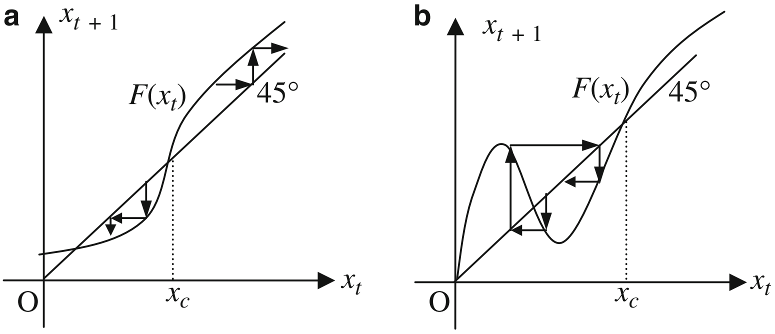
In Fig. 1a, $F(x)$, stays above the 45° line everywhere, hence the economy grows forever (as in the endogenous growth models). In Fig. 1b, for any x_0, the economy converges to x^* (as in the Solow growth model). In either case there is no poverty trap, since the long-run performance of the economy is independent of the initial condition, no matter how underdeveloped the economy is initially. (Confusion sometime occurs because a few authors use the term 'trap' to describe the situation depicted in Fig. 1b, in the sense that growth is not sustainable. However, this should more appropriately be called 'the limit to growth'. This limit is not caused by the initial poverty of the economy.)

In Fig. 2a and b, on the other hand, the long-run performance depends on the initial condition. When the economy starts above x_c, it will stay above x_c and may either grow forever or reach a higher stationary state. However, if it starts below x_c, it will be trapped forever below x_c. In this sense, both figures exhibit a poverty trap in its strong form. In Fig. 2a, the economy caught in the trap will converge to the low-level stationary state. In Fig. 2b, it will fluctuate below x_c. In both cases, the economy will remain poor only because it is poor. Thus, the poverty becomes its own cause. It is this self-perpetuating nature that sets 'the poverty trap' apart from 'the limit to growth'.

Both Fig. 2a and b project the very stark view that the economy can never escape from the poverty trap. This should not be taken too literally.



Poverty Traps, Fig. 1



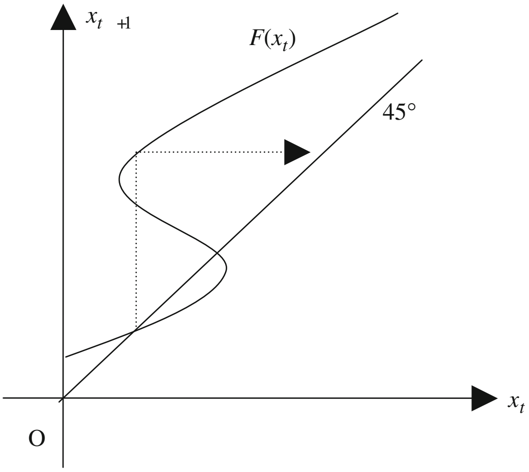
Poverty Traps, Fig. 2

The essential message of poverty traps is that poverty tends to persist, and that it is difficult, but not necessarily impossible, for the economy to escape from it. Poverty traps in their weak form are depicted in Fig. 3a and b. In Fig. 3a, the economy has to experience stagnation for long time as it travels through the 'narrow corridor' between $F(\cdot)$ and the 45° line, before eventually succeeding in taking off. In Fig. 3b, the economy may or may not manage to escape the trap after experiencing (possibly many) periods of volatility. For all practical purposes, the situations depicted in Fig. 2a and b and Fig. 3a and b are difficult to separate, but the message is the same: the self-perpetuating nature of poverty.

The above analysis can be extended in many directions. First, one could add stochastic shocks to the system, as $x_{t+1} = F(x_t, \xi_{t+1})$. Such shocks perturb the map, which may switch the graph back and forth between Figs. 2a (or 2b) and Figs. 3a (or 3b). This can be viewed as a jump in the state variable in the case of the additive shocks, $x_{t+1} = F(x_t) + \xi_{t+1}$. (For example, natural disasters, plagues and wars could cause the capital–labour ratio to jump up and down.) In the presence of such stochastic shocks, the economy may occasionally and recurrently escape or fall into the trap. Hence, the analysis has to be described in terms of the stochastic kernel; see Azariadis and Stachurski (2005) for a detailed discussion of stochastic poverty trap models.

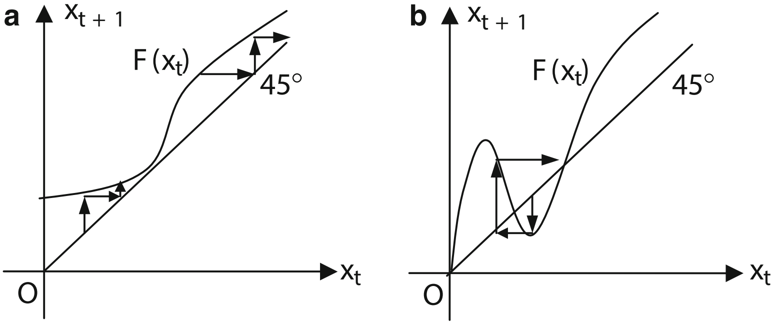
Second, the above analysis assumes that x_{t+1} is uniquely determined as a function of x_t. If the underlying economic models permit multiple equilibria, as often is the case with models of external economies and strategic complementarity, then $F(\cdot)$ becomes a correspondence, and the (deterministic) equilibrium path follows the difference inclusion, $x_{t+1} \in F(x_t)$. See Matsuyama (1997) for some examples. Figure 4 depicts one possibility, suggesting that the economy is stuck in a low-level stationary state, in part due to coordination failures. In this case, the economy could escape the poverty trap if it succeeded in coordinating on a higher equilibrium, as indicated by the dotted arrow. (If such coordination takes place through a realization of some coordination devices, 'sunspots', it can be viewed as a model of endogenous stochastic shocks.)

Third, the underlying economic model may imply that the law of motion be described in a multi-dimensional system. For example, the state space may be two-dimensional, $(x; q)$, where x is



Poverty Traps, Fig. 4

Poverty Traps, Fig. 3



the state (or backward-looking) variable, such as the capital stock, and q is the co-state (or forward-looking) variable, such as the asset price or consumption, and the law of motion is given by a two-dimensional difference equation, $(x_{t+1}, q_{t+1}) = F(x_t, q_t)$. In this case, for a given initial condition, x_0, the equilibrium condition may not uniquely pin down the initial value, q_0. That is, there may be multiple equilibrium paths, with self-fulfilling expectations, which suggests another way in which the economy may escape from the poverty trap; see Matsuyama (1991). Or the dimensionality of the state space may be equal to the number of industries in a multi-industry model, or to the number of countries in a multi-country world economy model. In such a high-dimensional system, one could encounter a much richer set of dynamics, where the long-run behaviour can depend on the initial condition in a much more complex manner.

Some Models of Poverty Trap

Many (dynamic) models of poverty traps have been proposed in the literature. The common feature of these models is the presence of some external economies or strategic complementarities that give rise to the circular causation. Here is a highly selective list.

Learning-by-Doing Externalities

The infant industry argument for protection (see Corden 1977, for a synopsis) is a classic example. When firms are inexperienced and unproductive, they cannot offer wages high enough to attract workers from other sectors, and hence are not able to accumulate experience. Temporary protection has been suggested as a way to break the vicious circle. Helping some industries accumulate experience to escape from a poverty trap, however, may end up pushing the economy into another poverty trap, as it could prevent other (new and possibly more promising) industries from growing. If the scope of productivity improvement in any industry is limited, then the only way of avoiding poverty traps and achieving sustainable growth is to keep the delicate balance so that production will shift constantly from one industry to another, as existing industries become mature and new industries are born; see Stokey (1988); Brezis et al. (1993); Matsuyama (2002).

Search Externalities

The difficulty of finding business partners can discourage many from entering an industry, which in turn makes it even harder for others to find business partners. See Diamond (1982).

Human Capital Externalities

Following the Lucas (1988) model of endogenous growth based on human capital accumulation, Azariadis and Drazen (1990) showed how it could lead to the existence of poverty traps, when human capital is subject to threshold externalities.

Market Size and Division of Labour

Adam Smith argued that 'the division of labour is limited by the extent of the market'. Young (1928) argued that the extent of the market is also limited by the division of labour. That is, economic growth can be achieved by means of greater specialization, which was formalized by Romer (1987) and others. Building on this body of work, Ciccone and Matsuyama (1996) showed how the economy can be caught in a poverty trap. The basic mechanism is that advanced technologies require the use of highly specialized equipment and producer services. In the underdeveloped economy, the limited availability of specialized inputs forces downstream industries to rely on less advanced technologies, which do not require the use of specialized inputs. This in turn leads to a small market size for specialized firms in upstream industries. Hence, the economy is caught in the vicious circle of limited market size and limited division of labour.

Financial Developments

In countries with limited opportunities to diversify risk, entrepreneurs are discouraged from making productive but risky investments. This in turn leads to a limited set of traded financial assets, which reduces the opportunity to diversify risk. See Saint-Paul (1992) and Acemoglu and Zilibotti (1997).

Low Wealth/Low Investment

When external finance is more costly than internal finance, a decline in borrower net worth leads to a higher investment distortion. In Bernanke and Gertler (1989), this leads to a decline in the investment, which in turn leads to a decline in the net worth of the next generation of entrepreneurs, hence generating persistence in the aggregate investment dynamics. In Matsuyama (2004), the same mechanism could make some (but not all) countries in the world caught in the vicious circle of low net worth–low investment. Matsuyama (2007) showed how the trap can sometimes take the form of greater volatility (as shown in Fig. 2b). In a set-up that allows for wealth distribution to evolve over time, Banerjee and Newman (1993) suggested that greater initial wealth inequality, to the extent that it increases the number of entrepreneurs rich enough to finance their investments, can lead to a higher aggregate investment, which in turn could help the poor in the long run, thereby breaking the vicious circle.

Demographic Trap

Nelson (1956) is among the first to argue that underdeveloped countries are caught in the vicious circle of high population growth and low per capita income. Becker et al. (1990) showed how the economy may be caught in the vicious circle of high fertility–low human capital. Basu (1999) and Doepke and Zilibotti (2005) discussed child labour traps. In Matsuyama (2000), intergenerational persistence of a high labour force participation rate by the elderly could lead to a poverty trap.

Contagious Social Norms

Tirole (1996) showed how corruption or other unethical behaviour can be contagious and persistent. He considered the setting where, in the presence of imperfect information, the reputation of a member of the group (say, a firm in the industry) depends not only on his own past behaviour, but also on the past behaviour of other group members. Then, when the group has the reputation of being dishonest, it would be difficult for the member to establish a reputation for honesty. This induces him to behave dishonestly, thereby contributing to the bad reputation of the group.

Modelling Inertia

Underdevelopment is often modelled as a Pareto-dominated equilibrium in a static game of strategic complementarities. Murphy et al. (1989) is the best-known example. By adding some inertia, which restricts the ability of the players to switch their strategies, one can convert virtually any static game of strategic complementarities into a dynamic model of poverty traps, where both the initial condition and expectations can play a role in determining the long-run performance of the economy. See the techniques developed by Matsuyama (1991) and Matsui and Matsuyama (1995).

Some Cautionary Remarks on Interpretations

The poverty trap is often interpreted as an explanation for *cross-country* income difference. As such, it is frequently viewed as an *alternative* to the models that attribute *cross-country* income difference to the *cross-country* difference in, say, TFP and/or investment distortions. This is a misinterpretation. First, the message of poverty trap models is the self-perpetuating nature of poverty. It suggests that the long-run performance of an economy could be much better if its initial condition were better. It does *not* mean that the cross-country difference in the long-run performance is due mostly to the difference in their initial conditions. Second, the notion of poverty trap does not contradict the observation that low income is often associated with low TFP and/or high investment distortions. Indeed, many poverty trap models attempt to explain the two-way causality between low-income and low TFP and/or high investment distortions. By endogenizing TFP and/or investment distortions, these poverty trap models go one step further than the models that treat these variables as exogenously given.

Many calls for foreign assistance for underdeveloped countries can be understood using the notion of poverty trap; see, for example, Sachs et al. (2004). Indeed, the poverty trap is often viewed as a powerful case for policy activism. However, one should be careful when using any particular model of the poverty trap to make policy proposals. It is important to keep in mind that each model of the poverty trap is designed to highlight one particular feedback mechanism behind the vicious circle. To this end, other sources of the poverty trap are deliberately assumed away. In reality, of course, many sources of the poverty trap are likely to coexist. If there is one important lesson from the literature reviewed above, it should be that there are hundreds of traps that the economy can fall into, and any policy intervention that attempts to pull the economy out of one trap may end up pushing it into another. As we know, any attempt to solve a problem can often become a source of another, even bigger problem. For more on this issue, see Matsuyama (1996), which discusses economic development as 'complex' coordination problems.

See Also

- ► Structural Change
- ► Symmetry Breaking

Bibliography

Acemoglu, D., and F. Zilibotti. 1997. Was Prometheus unbounded by chance?: Risk, diversification and growth. *Journal of Political Economy* 105: 709–751.

Azariadis, C., and A. Drazen. 1990. Thresholds externalities in economic development. *Quarterly Journal of Economics* 105: 501–526.

Azariadis, C., and J. Stachurski. 2005. Poverty traps. In *Handbook of economic growth*, ed. P. Aghion and S. Durlauf, vol. 1. Amsterdam: North-Holland.

Basu, K. 1999. Child labor: causes, consequences and cures, with remarks on international labor standards. *Journal of Economic Literature* 37: 1083–1119.

Banerjee, A., and A. Newman. 1993. Occupational choices and process of development. *Journal of Political Economy* 101: 274–298.

Becker, G., K. Murphy, and R. Tamura. 1990. Human capital, fertility, and economic growth. *Journal of Political Economy* 98: S12–S37.

Bernanke, B., and M. Gertler. 1989. Agency costs, net worth, and business fluctuations. *American Economic Review* 79: 14–31.

Brezis, E., P. Krugman, and D. Tsiddon. 1993. Leapfrogging in international competition: a theory of cycles in national technology leadership. *American Economic Review* 83: 1211–1219.

Ciccone, A., and K. Matsuyama. 1996. Start-up costs and pecuniary externalities in economic development. *Journal of Development Economics* 49: 33–57.

Corden, W. Max. 1977. *Trade policy and economic welfare*. Oxford: Clarendon Press.

Diamond, P. 1982. Aggregate demand management in search equilibrium. *Journal of Political Economy* 90: 881–894.

Doepke, M., and F. Zilibotti. 2005. The macroeconomics of child labor regulation. *American Economic Review* 95: 1492–1524.

Lucas, R. Jr. 1988. On the mechanics of economic development. *Journal of Monetary Economics* 22: 3–42.

Matsui, A., and K. Matsuyama. 1995. An approach to equilibrium selection. *Journal of Economic Theory* 65: 415–434.

Matsuyama, K. 1991. Increasing returns, industrialization, and indeterminacy of equilibrium. *Quarterly Journal of Economics* 106: 617–650.

Matsuyama, K. 1996. Economic development as coordination problems. In *The role of government in East Asian development: Comparative Institutional Analysis*, ed. M. Aoki, H. Kim, and M. Okuno-Fujiwara. New York: Oxford University Press.

Matsuyama, K. 1997. The 1996 Nakahara Lecture: complementarity, instability and multiplicity. *Japanese Economic Review* 48: 240–266.

Matsuyama, K. 2000. Economic development with endogenous retirement. CMS-EMS Discussion Paper No. 1237R. Northwestern University.

Matsuyama, K. 2002. The rise of mass consumption societies. *Journal of Political Economy* 110: 1035–1070.

Matsuyama, K. 2004. Financial market globalization, symmetry-breaking, and endogenous inequality of nations. *Econometrica* 72: 853–884.

Matsuyama, K. 2007. Credit traps and credit cycles. *American Economic Review* 97: 503–516.

Murphy, K., A. Shleifer, and R. Vishny. 1989. Industrialization and the big push. *Journal of Political Economy* 97: 1003–1026.

Nelson, R. 1956. A theory of the low level equilibrium trap in underdeveloped economies. *American Economic Review* 46: 894–908.

Romer, P. 1987. Growth based on increasing returns due to specialization. *American Economic Review* 77: 56–62.

Sachs, J., et al. 2004. Ending Africa's poverty trap. *Brookings Papers on Economic Activity* 1(2004): 117–240.

Saint-Paul, G. 1992. Technology choice, financial markets and economic development. *European Economic Review* 36: 763–781.

Stokey, N. 1988. Learning-By-Doing and the introduction of new goods. *Journal of Political Economy* 96: 701–717.

Tirole, J. 1996. A theory of collection reputations (with applications to the persistence of corruption and to firm quality). *Review of Economic Studies* 63: 1–22.

Young, A. 1928. Increasing returns and economic progress. *Economic Journal* 38: 527–542.

Power

Samuel Bowles and Herbert Gintis

Abstract

We consider the exercise of power in competitive markets for goods, labour and credit. We offer a definition of power and show that if contracts are incomplete it may be exercised either in Pareto-improving ways or to the disadvantage of those without power. Contrasting conceptions of power including bargaining power, market power, and consumer sovereignty are considered. Because the exercise of power may alter prices and other aspects of exchanges, abstracting from power may miss essential aspects of an economy. The political aspect of private exchanges challenges conventional ideas about the appropriate roles of market and political competition in ensuring the efficiency and accountability of economic decisions.

Keywords

Bargaining power; Coase, R.H.; Consumer sovereignty; Firm, theory of; Incomplete contracts; Labour market contracts; Labour market search; Market power; Monopolistic competition; Nash equilibrium; Pareto efficiency; Power; Principal and agent; Purchasing power; Rent; Reservation wage; Sanctions; Short-side power; Technical efficiency

JEL Classifications

D6

Power is exercised in the competitive markets for goods, labour and credit. We consider this aspect of economic power, setting aside the widely recognized exercise of power by members of governments and other coercive bodies and the influence of economic groups on governmental policy.

Background

'An economic transaction is a solved political problem.…', wrote Abba Lerner (1972, p. 259)'. . .economics has gained the title Queen of the Social Sciences by choosing solved political problems as its domain'. Prior to the development of modern contract theory, the standard approach to power among economists was aptly summed up by Paul Samuelson (1957, p. 894), 'Remember that in a perfectly competitive market, it really does not matter who hires whom; so have labor hire capital'. As if responding to Samuelson, John Kenneth Galbraith (1967, p. 47), chided economists for not having asked 'why power is associated with some factors [of production] and not with others?' But with some notable exceptions (for example, Zeuthen 1930; Shapley and Shubik 1967; Samuels 1973; Lindblom 1977; Basu 1986; Takada 1995; Hirshleifer 1991; Chichilnisky and Heal 1984; Lundberg and Pollak 1994; Rotemberg 1993; Pagano 1999; Bardhan 2005; Aghion and Tirole 1997) economists have treated power as the concern of other disciplines and extraneous to economic explanation. The term does not appear among the 1,300 or so index entries of the leading graduate microeconomics text (Mas-Colell et al. 1995).

The reason is that Samuelson's claim is true in the Walrasian model: if contracts are complete, 'hiring' simply means 'buying'.'What does it mean', Oliver Hart (1995) asked, 'to put someone "in charge" of an action or decision if all actions can be specified in a contract?' But as an empirical matter, as Marx (1867), Coase (1937), Simon (1951) and others have stressed, the firm is a political institution in the sense that some members of the firm routinely give commands while others are constrained by the threat of sanctions to obey. To say that the manager has the right to decide what the worker will do means only that he has the legitimate authority to do this, not the power to secure compliance. Given that in a liberal economy management is sharply restricted in the kinds of punishment they can inflict, and given that the employee is free to leave, the fact that orders are typically obeyed is a puzzle. Why, in Coase's initial formulation, is the command of the manager (to move ‘from department Y to department X’) obeyed (Coase 1937)?

Noticing the lack of a good answer, Alchian and Demsetz (1972) challenged the Coasean idea that the firm is a mini ‘command economy’, suggesting that the employment contract is no different in this respect from other contracts.

> The firm. . .has no power of fiat, no authority, no disciplinary action any different in the slightest degree from ordinary market contracting between any two people. . .Wherein then is the relationship between a grocer and his employee different from that between a grocer and his customer? (1972, p. 777)

Hart (1989, p. 1771) offered the following response to Alchian and Demsetz:

> . . .the reason that an employee is likely to be more responsive to what his employer wants than a grocer is that the employer. . .can deprive the employee of the assets he works with and hire another employee to work with these assets, while the customer can only deprive the grocer of his customer and as long as the customer is small, it is presumably not very difficult for the grocer to find another customer.

Hart motivates the difference between the grocer and the employer by the assumption that the employee needs access not just to a job (and hence *some* assets) but *to this particular employer’s assets*. This might be the case due to a complementarity between the two (the employee may have made an investment in training which is of value only when combined with this particular asset, for example). Other less obvious (and probably more important) examples come to mind. Excluding an employee from access to a particular asset may require the employee to relocate, disrupting family and friendships. The loss of a job may also harm the employee’s reputation.

While transaction-specific investments of this type undoubtedly explain some authority relationships – in company towns, and for some professional jobs and managers, for example – the explanation seems insufficiently general to provide an adequate explanation of the entire authority structure of the firm, especially in large urban labour markets and for non-professional employees. We thus need a complementary explanation based on the fact that the employee excluded from access to *her current employer’s asset* may not find access to *any asset* even in a competitive economy in which transaction-specific assets are absent. This will require clarity about what we mean by power.

Power as a Political Means to Gain Economic Advantage in Private Exchange

Because of its close connection to value-laden words such as ‘coercion’ and ‘freedom’ the term itself has proven to be controversial among philosophers and political theorists (Nozick 1969; Lukes 1974; Bachrach and Baratz 1962; Barry 1976; Taylor 1982). Nonetheless, common usage suggests several characteristics that must be present when power is said to be exercised. First, power is *interpersonal*, an aspect of a relationship among people, not a characteristic of a solitary individual. Second, the exercise of power involves the *threat and use of sanctions*. Indeed, many political theorists regard sanctions as the defining characteristic of power.

Lasswell and Kaplan (1950, p. 75) make the use of ‘severe sanctions. . .to sustain a policy against opposition’ a defining characteristic of a power relationship, and Parsons (1967, p. 308) regards ‘the presumption of enforcement by negative sanctions in the case of recalcitrance’ a necessary condition for the exercise of power. Third, the concept of power should be *normatively indeterminate*, allowing for Pareto-improving outcomes (as has been stressed by students of power from Hobbes to Parsons), but also susceptible to abuse in ways that harm others in violation of ethical principles. Finally, power must be *sustainable as a Nash equilibrium* of an appropriately defined game. Power may be exercised in disequilibrium situations, of course, but, as an enduring aspect of social structure, it should be a characteristic of an equilibrium.

The following sufficient condition for the exercise of power captures these four desiderata: *For B to have power over A, it sufficient that, by imposing or threatening to impose sanctions* *on A, B is capable of affecting A's actions in ways that further B's interests, while A lacks this capacity with respect to B* (Bowles and Gintis 1992).

The fact that sanctions are essential to the exercise of power in our sense makes it distinct from other means of influencing the behaviour of others that may operate even in the complete absence of strategic interaction, as in a Walrasian market setting. Consider, for example, the standard definition due to Robert Dahl (1957, pp. 202–3): 'A has power over B to the extent that he can get B to do something that B would not otherwise do.' But one can affect the behaviour of another in ways that do not involve power in the usual sense of that term. If we buy a commodity, there will be a whole series of market effects through the economy which entail others doing things they would not otherwise have done. But to say that our purchase of bread is an exercise of power over some unknown wheat farmer with whom we do not interact strategically is to expand the concept of power beyond recognition. By making the threat of sanctions a necessary aspect of power we also exclude forms of interpersonal influence such as persuasion and the provision of information.

Short-Side Power in Labour, Credit and Goods Markets

The power that may be exercised by an economic actor depends on the actor's position in the institutions of society. Power may be exercised by economic actors who are on the short side of a non-clearing market, namely, the side of the market on which the number of desired transactions is less, that is, employers in a labour market with unemployment, lenders in a loan market with borrowers facing credit constraints, and so on. Because those holding power in these cases are those on the short side of the market, we term this 'short-side power'. This clarifies the difference between the employer and the grocer in Hart's response to Alchian and Demsetz: the sanctions imposed on the employee by depriving him of access to the capital good are severe because, in a labour market with perpetual excess supply of labour, finding another job will be difficult, while the costs imposed on the grocer by the departing customer are negligible or zero. The reason why the consumer, in switching to another seller, does not impose a sanction on the grocer is that the grocer (in competitive equilibrium) was maximizing profits by selecting a level of sales that equates marginal cost to the exogenously given price, and, this being the case, a small variation in sales has only a second-order effect on profits.

Let us check to see that this conception of power applies to the employment relationship in which transaction specificity is absent. We know that in a standard labour-discipline model (Gintis 1976; Shapiro and Stiglitz 1984; Bowles 1985), in equilibrium the worker receives a rent: the present value of the job exceeds her next-best alternative (job search) and, because she fears losing his job, she works harder than she would have in the absence of the employer's incentive strategy. These results together imply that the employer has caused the worker to act in the employer's interest by credibly threatening to sanction the worker. The employee lacks this capacity with respect to the employer for, were the employee to threaten the employer with a sanction should he not raise the wage (to damage his machinery or beat him up or simply to work less hard), the threat would not be credible. The employer would simply refuse to respond, knowing that it would not be in the interest of the employee to carry out the threat.

Note that the exercise of power allows a Pareto improvement over a counterfactual condition in which power cannot be exercised, namely, that the worker is hired at her reservation wage and works at the reservation effort level. This follows directly as we know from the fact that the worker receives an equilibrium rent at the wage offered by the employer. Both expected worker lifetime utility and firm profits are higher in equilibrium (with power being exercised) than at the (power-absent) reservation position. This is yet another example of a situation in which the exercise of power helps to address coordination failures, albeit sometimes with objectionable consequences those without power. An example from Bowles (2004) follows.

Suppose the employer determines (in addition to the wage) some aspect of the job affecting workplace amenities, including not only such innocuous things as the quality of the music on the office sound system but also management practices affecting the employee's dignity, such as not being subjected to racial insults, sexual harassment or other on-the-job indignities. If the firm sets these amenities to maximize profits, it follows that the employer can inflict first-order costs on the worker (by reducing the amenity a small amount) at second-order cost to himself (the costs are second-order because due to profit maximization the derivative of profits with respect to the level of amenities is zero). Thus the competitive equilibrium in an employment relationship gives the employer the capacity not only to exercise power to attenuate coordination problems but also to exercise power arbitrarily, that is, to inflict costs on another at virtually no cost to himself. When this power is exercised in unethical ways it may be termed coercive.

Thus the strategic interaction between the employer and employee allows the exercise of power in a manner conforming to the four desiderata outlined above: sanctions are credibly threatened (and used) in a strategic interaction describing a Nash equilibrium, and the resulting exercise of power is Pareto-improving over a reasonable counterfactual but may also be used coercively.

It is easy to check that power in the sense defined may be exercised in the standard principal–agent model of the credit market as well. The lender offers the borrower terms that are preferred to the borrower's reservation position, promising to make additional loans in the future if the borrower repays the loan. In this contingent renewal model, the borrower pursues a less risky strategy than would have been the case had the lender not offered a rent. Where the borrower's participation constraint holds as an equality, power in the sense defined cannot be exercised for the simple reason that the borrower is indifferent between the current transaction and the next-best alternative, so the only sanction permitted in a liberal economy – termination of the contract – has no force.

Short-side power may be contrasted with the 'markets and hierarchies' approach pioneered by Oliver Williamson (1985). Rather than seeing firms simply as 'islands of conscious power in this ocean of unconscious cooperation', in Robertson's (1923, p. 85) apt words, the incomplete contracts approach traces the exercise of power to both the structure of markets and the structure of firms. The firm is an important venue in which power is exercised, but, as the credit market model makes clear, power may be exercised in the absence of firms or indeed any organizational structure whatsoever. Short-side power is exercised *in* markets, not simply outside markets or despite markets.

Wealth, Power and 'Consumer Sovereignty'

Thus an agent's location in the economic structure of a society – on the short side of a non clearing market – may make it possible for him to exercise power over others. How are agents assigned to these positions of short-side power? Given that employing others requires capital and that borrowing substantial amounts typically requires that the borrower have sufficient wealth to invest in the project or to provide collateral, an important determinant of an individual's assignment to a position of short-side power is the individual's wealth. The wealthy may exercise power over those to whom they lend, who in turn may exercise power over those (managers or other employees) whom they hire. As a result, power cascades downward from the loan market to the market for managers to the market for non-managerial employees (Bowles 2004).

A less obvious case concerns the power of the consumer, sometimes summarized by the term 'consumer sovereignty'. Consider a principal–agent model involving difficult-to-measure product quality (Klein and Leffler 1981; Gintis 1976). In equilibrium, the buyer pays the seller a price exceeding the seller's next-best alternative and promises continued purchases contingent on the seller providing high-quality goods. The seller's prospect of losing the resulting rent conferred by the buyer induces the seller to provide higher quality than would have been provided in the absence of the threatened sanction. Thus the buyer has exercised power over the seller in the sense just defined.

As the example suggests, buyers may exercise power over sellers whenever the buyer's threat to switch to an alternative seller is credible and inflicts a cost on the seller. Consider two monopolistically competitive sellers (that is, firms facing downward-sloping demand functions) and a consumer who is indifferent between purchasing from one or the other. Both sellers have chosen a level of output to maximize profits, setting marginal cost equal to marginal revenue (which is less than the price because the demand curve is downward sloping). For both sellers, price thus exceeds marginal cost, and as a result the consumer's choice confers a rent on one and deprives the other of the rent. The reader may wonder how the rent can arise if the firm has chosen the output level to maximize profits, each setting the derivative of profits with respect to sales equal to zero. But the buyer's switch from one to the other seller is not a movement *along* a demand function (the basis of the firm's output choice), but rather is a horizontal *shift* in the demand function (inwards for the firm the consumer rejected, outwards for the firm to which he switched). As a result of the switch, for the fortunate firm it is profit maximizing to sell one more unit at the going price.

Ironically, the idealized Walrasian conditions under which consumer sovereignty is said to hold give the consumer no power in the sense defined here, while deviations from the canonical competitive assumption that price equals marginal cost (because firms face downward sloping demand functions) create an environment in which the consumer may exercise power. Of course, the strategic position of the consumer as one of many principals facing a single agent is quite unlike that of the employer facing many potential employees or the lender facing many potential borrowers. As Hart observed about the consumer and the grocer, a single consumer will not generally be in a position to command the supplier to improve the product quality and expect the supplier to obey. The power of consumers is thus limited by the difficulties the many principals face in acting in a coordinated fashion.

Non-clearing Markets and Inefficient Competitive Equilibria

Where power is exercised by a principal who confers a rent on an agent and monitors the agent's actions – as in the markets for labour, credit and goods just analysed – the equilibrium allocation will generally be neither Pareto-efficient nor technically efficient. The reason for the first is that the principal is constrained not by the agent's reservation utility but by the agent's best-response function. As a result, small changes in the instruments controlled by the principal – the wage, the rate of interest or the price – incur only second-order costs or benefits for the principal but first-order benefits and costs for the agent. For the actions controlled by the agent the reverse is true. Therefore, there must exist some set of small variations away from the equilibrium allocation that improve the utility of both principal and agent. A labour market example of such a Pareto improvement is a small increase in the wage accompanied by a small increase in worker effort.

The allocation will be technically inefficient because the principal chooses the enforcement strategy with respect to the private costs (the costs of both the rent conferred on the agent and the monitoring) while there is no social cost associated with the rent (because, unlike the monitoring costs, it is a pure transfer and is not resource using). From the equilibrium allocation, therefore, there must exist a technical efficiency-improving increase in the agent's rent and a reduction in monitoring.

Exploiting these potential efficiency gains requires changes in the information and incentive structure of the interaction, for example by making the agent the residual claimant on his or her non-contractible actions, if this is possible.

The three cases for which we have analysed the exercise of power – by the buyer over the seller, the lender over the borrower, and the employer over the employee – are members of a generic class of power relationships which are sustainable in the equilibrium of a system of voluntary competitive exchanges. In all three, those with power are transacting with agents who receive rents and hence are not indifferent between the current transaction and their next-best alternative. This being the case, there must exist other identical agents who are quantity constrained, namely, the unemployed, those excluded from the loan market or restricted in the amount they can borrow, and sellers who fail to make a sale. For this situation to characterize an equilibrium it must be that markets do not clear, which, as we have seen will be the case.

Power as we have defined it can be exercised in other ways, even when markets clear. An interesting (if perhaps not empirically important) example is provided by the case of optimal job fees, in which the fee eliminates the job rent *ex ante* so the market clears, the worker being indifferent between taking the job and paying the fee or not. But an *ex post* rent nonetheless exists, giving the employer the ability to sanction the employee. A job fee of this type is a pure case of an employee's transaction-specific investment, and the basis of the power of the employer in this case is an example of Hart's reasoning, above.

All three of those exercising power in the above examples – buyer, lender, employer – have in common that the party that contributes money to the transaction – the buyer's purchase price, the lender's loan, the employer's wage offer – is the one exercising power. This may seem an analytical foundation for the familiar adage that 'money talks', but the conclusion is misleading. Recall that in the centrally planned Communist economies it was generally the case that consumer durables (and many other consumer goods) sold below market-clearing prices. The resulting excess demand was allocated through a process of queuing and by other means (Kornai 1980). In this case the producers (sellers) were on the short side of the market, and those bringing money to the transaction – the buyers – were the long-siders, some of whom failed to make a trade. The notorious inferiority in the quality of consumer goods in centrally planned economies to those in capitalist economies may be explained in part by the fact that consumers were long-siders in the former and shortsiders in the latter. Or, to put it more graphically, one reason why Fords were better cars than their Cold War era Russian equivalents is that in Russia customers waited in line to purchase Volgas while in the United States Ford salesmen lined up to sell customers cars. Another reason is that in the United States workers waited in line to get jobs at Ford.

Other Conceptions of Power

Other uses of the term 'power' are common in economics. (We do not address the concept of 'coalitional power' advanced by Shapley and Shubik 1954, as it has found application primarily in the analysis of committees voting and other arenas addressed by political scientists.) 'Purchasing power' is just another word for the position of one's budget constraint (or wealth), and it does not concern the exercise of sanctions or indeed any strategic interaction at all. 'Market power' arises in thin markets in which an actor can benefit by varying a price. In the standard monopolistic competition case the seller is said to have market power. The seller is less constrained in the sense that he faces a downward sloping demand rather than horizontal demand function, while the consumer is more constrained in that there may be less choice among suppliers. But we have just seen that in this case the consumer who switches from one seller to another confers a rent on his favoured firm. (This why Ford salesmen line up to sell you cars.) Thus, if the buyer can credibly threaten to withdraw the rent he may be able to exercise short-side power over the seller. It thus is not clear how to reconcile usual notions of power – the use of sanctions to gain advantage – with the statement that the monopolist has power over the consumer.

Finally, there is 'bargaining power', typically meaning the share of the joint surplus which a party gains in a bargain (Binmore et al. 1986). Reflecting this usage, the exponents used in the

'Nash product' to solve the generalized Nash bargaining model are said to refer to the bargaining power of the two parties. Used this way, bargaining power refers to outcomes – to how much advantage one may gain – rather than to any particular means of attaining it (for example by threatening a sanction). If the bargaining problem is embedded in an ongoing interaction, then bargaining power and short-side power appear not only unrelated but even opposed. In the competitive equilibrium of the standard principal–agent model of the labour market, for example, the principal receives his reservation return (given by the zero profit condition) while the agent receives a rent. Therefore, the bargaining-power perspective would say that the employee has *all* the bargaining power. But the short-side power perspective would conclude that, far from a sign that the employee is powerful, the rent conferred on the employee as a profit-maximizing choice of the employer is the reason why the employer has power over the employee. The employee receives the rent because his services cannot be costlessly contracted for, and the employer profits in this case by paying to exercise power over the employee.

The fact that the exercise of power is ubiquitous in private exchange shows that it is mistaken to think of society as composed of a political sphere, meaning governments and other bodies with formal powers of coercion, and a private economic sphere in which the exercise of power is absent. The rejection of this public–private division raises important issues concerning the appropriate scope of for democratic political competition (in addition to market competition) as a guarantor of accountability in the economy (Dahl 1977; Bowles and Gintis 1993).

See Also

- ▶ Bargaining
- ▶ Labour Market Institutions

Acknowledgment We are grateful to Maria Olivella Rizza, Ceren Soylu, Elisabeth Wood and Peyton Young for comments, and to the Behavioral Sciences Program of the Santa Fe Institute, the University of Siena and the MacArthur Foundation for financial support.

Bibliography

Aghion, P., and J. Tirole. 1997. Formal and real authority in organizations. *Journal of Political Economy* 105: 1–29.

Alchian, A.A., and H. Demsetz. 1972. Production, information costs, and economic organization. *American Economic Review* 62: 777–795.

Bachrach, P., and M. Baratz. 1962. The two faces of power. *American Political Science Review* 56: 947–952.

Bardhan, P. 2005. *Scarcity, conflicts and cooperation*. Cambridge, MA: MIT Press.

Barry, B., eds. 1976. *Power and political theory: Some European perspectives*. New York: John Wiley.

Basu, K. 1986. One kind of power. *Oxford Economic Papers* 38: 259–282.

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

Bowles, S. 1985. The production process in a competitive economy: Walrasian, neo-Hobbesian, and Marxian models. *American Economic Review* 75: 16–36.

Bowles, S. 2004. *Microeconomics: Behavior, institutions, and evolution*. Princeton: Princeton University Press.

Bowles, S., and H. Gintis. 1992. Power and wealth in a competitive capitalist economy. *Philosophy and Public Affairs* 21: 324–353.

Bowles, S., and H. Gintis. 1993. A political and economic case for the democratic enterprise. *Economics and Philosophy* 9: 75–100.

Chichilnisky, G., and G. Heal. 1984. Patterns of power: Bargaining and incentives in two-person games. *Journal of Public Economics* 23: 333–349.

Coase, R.H. 1937. The nature of the firm. *Economica* 4: 386–405.

Dahl, R.A. 1957. The concept of power. *Behavioral Science* 2: 201–215.

Dahl, R.A. 1977. On removing certain impediments to democracy in the United States. *Political Science Quarterly* 92: 1–20.

Galbraith, J.K. 1967. *The new industrial state*. Boston: Houghton Mifflin.

Gintis, H. 1976. The nature of the labor exchange and the theory of capitalist production. *Review of Radical Political Economics* 8(2): 36–54.

Gintis, H. 1989. The power to switch: On the political economy of consumer sovereignty. In *Unconventional wisdom: Essays in honor of John Kenneth Galbraith*, ed. S. Bowles, R. Edwards, and W.G. Shepherd. New York: Houghton-Mifflin.

Hart, O. 1989. An economist's perspective on the theory of the firm. *Columbia Law Review* 89: 1757–1774.

Hart, O. 1995. *Firms, contracts, and financial structure*. Oxford: Clarendon Press.

Hirshleifer, J. 1991. The paradox of power. *Economics and Politics* 3: 177–200.

Klein, B., and K. Leffler. 1981. The role of market forces in assuring contractual performance. *Journal of Political Economy* 89: 615–641.

Kornai, J. 1980. *Economics of shortage*. Amsterdam: North-Holland.

Lasswell, H., and A. Kaplan. 1950. *Power and society: A framework for political enquiry*. New Haven: Yale University Press.

Lerner, A. 1972. The economics and politics of consumer sovereignty. *American Economic Review* 62: 258–266.

Lindblom, C.E. 1977. *Politics and markets: The world's political-economic systems*. New York: Basic Books.

Lukes, S. 1974. *Power: A radical view*. London: Macmillan.

Lundberg, S., and R. Pollak. 1994. Noncooperative bargaining models of marriage. *American Economic Review* 84: 132–137.

Marx, K. 1867. *Capital: A critique of political economy. I: The process of capitalist production*. New York: International Publishers.

Mas-Colell, A., M.D. Whinston, and J.R. Green. 1995. *Microeconomic theory*. New York: Oxford University Press.

Nozick, R. 1969. Coercion. In *Philosophy, science and method*, ed. S. Morgenbesser, P. Suppes, and M. White. New York: St. Martins Press.

Pagano, U. 1999. Is power an economic good? Notes on social scarcity and the economics of positional goods. In *The politics and economics of power*, ed. S. Bowles, M. Franzini, and U. Pagano. London: Routledge.

Parsons, T. 1967. On the concept of political power. In *Sociological theory and modern society*, ed. T. Parsons. New York: Free Press.

Robertson, D. 1923. *The control of industry*. Cambridge: Cambridge University Press.

Rotemberg, J.J. 1993. Power in profit-maximizing organizations. *Journal of Economics & Management Strategy* 2: 165–198.

Samuels, W.J. 1973. The economy as a system of power and its legal bases: The legal economics of Robert Lee Hale. *University of Miami Law Review* 27: 262–371.

Samuelson, P. 1957. Wages and interest: A modern dissection of Marxian economics. *American Economic Review* 47: 884–921.

Shapiro, C., and J. Stiglitz. 1984. Unemployment as a worker discipline device. *American Economic Review* 74: 433–444.

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. 1967. Ownership and the production function. *Quarterly Journal of Economics* 81: 88–111.

Simon, H. 1951. A formal theory of the employment relation. *Econometrica* 19: 293–305.

Takada, Y. 1995. *Power theory of economics*. New York: St. Martin's Press.

Taylor, M. 1982. *Community, anarchy, and liberty*. Cambridge: Cambridge University Press.

Williamson, O.E. 1985. *The economic institutions of capitalism*. New York: Free Press.

Zeuthen, F. 1930. Economic warfare. In *Problems of monopoly and economic warfare*. New York: A.-M. Kelly 1968.

Power Laws

Xavier Gabaix

Abstract

A power law is the form taken by a remarkable number of regularities in economics, and is a relation of the type $Y = kX^{\alpha}$, where Y and X are variables of interest, α is called the power law exponent, and k is a constant. Many economic laws take the form of power laws, in particular macroeconomic scaling laws, the distribution of income, wealth, size of cities and firms, and the distribution of financial variables such as returns and trading volume. This article surveys the empirical evidence and the theoretical explanations for the occurrence of power laws.

Keywords

Cities; GARCH effects; Gibrat's law; Matching; Networks; Pareto laws; Power laws; Proportional random growth; Quantity theory of money; Scaling laws; Stock market volatility; Stylized facts; Superstars, economics of; Trading volume; Universality; Urban economics; Zipf's law

JEL Classifications

D85

A power law (PL), also known as a scaling law, is the form taken by a remarkable number of regularities or 'laws' in economics, and is a relation of the type $Y = kX^{\alpha}$, where Y and X are variables of interest, α is called the power law exponent, and k is a typically unremarkable constant.

A special type is the distributional PL, also called a Pareto law. For instance, the probability that a firm has more than x employees is proportional to $1/x^{\zeta}$, for some positive number ζ: $P(S > x) = k/x^{\zeta}$, for some k, at least in the upper tail or most of it. The exponent ζ is independent of the units in which the law is expressed. A special case is Zipf's law, which is a Pareto law with $\zeta \simeq 1$.

Understanding what gives rise to the scaling law, and explaining the precise value of the exponent (for example, why it is equal to 1 rather than any other number) is a challenge that has fascinated successive generations. Schumpeter (1949, p. 155) wrote: ‘Few if any economists seem to have realized the possibilities that such invariants hold for the future of our science. In particular, nobody seems to have realized that the hunt for, and the interpretation of, invariants of this type might lay the foundations for an entirely novel type of theory.’ Champernowne (1953) and Simon (1955) made great strides towards realizing Schumpeter’s vision, and the quest continues.

Power laws are also of great interest outside of economics. Understanding PLs is a large part of the theory of critical phenomena, in which many materials behave identically around phase transitions – a phenomenon physicists call ‘universality’, and which is still only partially understood. Power laws have proven useful for describing and understanding networks. Biology has also many scaling regularities; for example, the daily energy intake of an animal of mass M is proportional to the $M^{3/4}$. This regularity was explained (Brown and West 2000) via simple physical reasoning, which eschews the need to talk about the feathers and the hair of animals. Simpler and deeper principles underlie the regularities instead. The same holds for economic laws. Power laws give the hope of robust, detail-independent economic laws.

Theory: Forces That Generate Power Laws

Proportional Random Growth

Getting a power law To explain distributional PLs, a central mechanism is proportional random growth (Sornette 2001). The process was developed in economics by Champernowne (1953) and Simon (1955). Things are more tractable in continuous time (see Gabaix 1999).

Take the example of cities in an economy with a constant number of cities and a fixed total population. When the system grows, the same reasoning applies after normalization – S is the normalized size of a city, for example as a multiple of the median city population. Suppose that each city i has a population S_t^i and, between t and $t+1$, increases by a growth rate γ_{t+1}^i:

$$\mathrm{S}_{t+1}^i = \gamma_{t+1}^i \mathrm{S}_t^i \qquad (1)$$

and suppose that the γ_{t+1}^i are identically and independently distributed, with density $f(\gamma)$, at least in the upper tail. Call $G_t(x) = P\left(\mathrm{S}_t^i > x\right)$ the counter-cumulative distribution function. The equation of motion of G is:

$$\begin{aligned} G_{t+1}(x) &= P\left(S_{t+1}^i > x\right) = P\left(S_t^i > x/\gamma_{it}\right) \\ &= E[G_t\left(x/\gamma_{it}\right)]. \end{aligned}$$

Hence:

$$G_{t+1}(S) = \int_0^\infty G_t\left(\frac{S}{\gamma}\right) f(\gamma) d\gamma.$$

Its steady state distribution G, if it exists, satisfies

$$G(S) = \int_0^\infty G\left(\frac{S}{\gamma}\right) f(\gamma) d\gamma. \qquad (2)$$

One can try the functional form $G(S) = a/S^\zeta$, where a is a constant. Plugging it in Eq. (2) gives: $1 = \int_0^\infty \gamma^\zeta f(\gamma) d\gamma$, that is

$$E\left[\gamma^\zeta\right] = 1. \qquad (3)$$

The steady state distribution is (in the upper tail) Pareto, with an exponent ζ that satisfies Eq. (3).

To make sure that the steady state distribution exists, one needs some friction, for example a force that prevents small cities from becoming too small.

Getting a Zipf’s law We see that proportional random growth leads to a PL. Why should the exponent $\zeta = 1$ appear in so many economic systems? An answer is the following (see Gabaix

1999; Luttmer 2007; Rossi-Hansberg and Wright 2007). Suppose that the random growth process Eq. (1) holds through most of the distribution, and that the system has constant size. Then, $E[S_{t+1}] = E[\gamma]E[S_t]$. As the system has constant size, then we need $E[S_{t+1}] = E[S_t]$, hence $E[\gamma]= 1$. That means that $\zeta = 1$ is a solution of Eq. (3). In other words, to get Zipf's law we need a random growth process with small frictions.

In sum, proportional random growth with frictions leads to PLs, and proportional random growth with small frictions leads to a special type of PL, namely Zipf's law.

Inheritance Via Algebraic Transformation

Power laws have excellent inheritance and aggregation properties. The property of being distributed according to a PL is conserved under addition, multiplication, power transformation, min, and max. The general rule is that, when we combine two PL variables, the fatter-tailed (that is, the one with the smaller exponent) dominates. Call ζ_X the PL exponent of X, with $\zeta_X = +\infty$ if X is thinner than any PL, for example is a Gaussian. For X and Y independent random variables, and $\beta > 0$ a constant, we have: $\zeta_{X+Y} = \zeta_{X\cdot Y} = \zeta_{\max(X;Y)} = \min(\zeta_X\,;\zeta_Y)$, $\zeta_{\min(X;Y)} = \zeta_X + \zeta_Y$, $\zeta_{\alpha X} = \zeta_X, \zeta_X{}^{\alpha}/\zeta_X = \alpha$ (see Jessen and Mikosch 2006). Those properties generate new PLs from old ones. For instance, if mutual funds are PL distributed, then many of their actions (for example, trading volumes, or the price movements they create) will be PL distributed (Gabaix et al. 2006).

Equilibrium Economic Mechanisms

Optimization with PL objective function The early example is the Allais–Baumol–Tobin model of demand for money (see also Mulligan and Shleifer 2005; Gabaix et al. 2003). Costs and benefits are power functions of the variables of interest, so that maximization also yields a PL – there, money demand is proportional to the interest rate to the power $-1/2$. PL in, PL out.

Matching talents in the upper tail Another way to generate PLs is in matching the talent of individuals with large firms or audiences. For instance, Gabaix and Landier (2008) study the market for executives. They derive that, in the upper tail of all well-behaved distributions, if $T(x)$ is the talent of an individual in the x upper quantile, then $T'(x)$ is approximately a power function x^{α}. As a result, the competitive matching process generates a PL relation between CEO pay and firm size, and a PL of the pay distribution. Huge differences in pay reward minuscule differences in talent. The PL form of T' is likely to be useful in other superstars markets.

Empirics: The Main Power Laws of Economics

Old Macroeconomic Scaling Laws

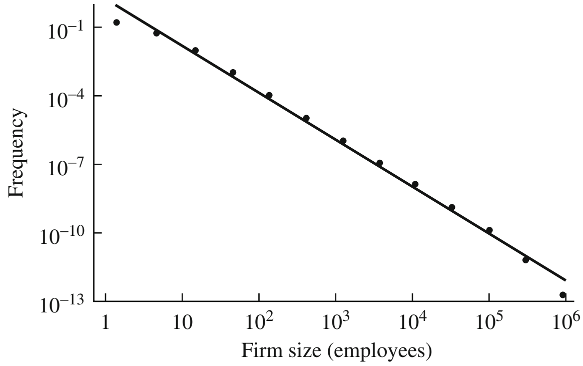
The first quantitative law of economics is probably the quantity theory of money, which, not coincidentally, is a scaling relation. It states that the price level P is proportional to the mass of money in circulation M, divided by the gross domestic product Y, times a pre-factor V: $P = VM\ /Y$. If the money supply doubles while GDP remains constant, prices double – a nice scaling law, relevant to policy.

More modern, we have Kaldor's stylized facts on economic growth: with K the capital stock, Y GDP, L population, r the interest rate, K/Y, wL/Y, and r, are roughly constant across time and countries. Explaining these facts led Solow to his growth model.

Reasonably Old and Well-Established Laws

Income and wealth The first PL is the Pareto law of income or wealth, which states that the tail distribution of income (or, respectively, wealth), is PL. The tail exponent of income seems to vary between 1.5 and 3, while the tail exponent of wealth is more stable. While, starting with Champernowne (1953), many models have been proposed to explain it (mainly along the lines of random growth), it is intriguingly unclear why the exponent is rather stable across economies.

Firm sizes The bulk of the distribution of firm sizes is well described by a Zipf's law (Fig. 1). This severely constrains models of firm growth,



Power Laws, Fig. 1 Note: Log frequency ln $f(S)$ vs. log size ln S of US firm sizes for 1997. OLS fit gives a slope of $1 + \zeta = 2{:}059$(s.e. $= 0.054$; $R^2 = 0.992$). This corresponds to a frequency $f(S) = kS^{-2{:}059}$, that is, a power law distribution with exponent $\zeta = 1{:}059$. Indeed, if $P(\text{Size} > S) = kS^{-}\zeta$ the density is $f(S) = k\zeta S^{-(\zeta+1)}$. This is very close to Zipf's law, which says that $\zeta = 1$ (Source: Reprinted with permission in Fig. 1 from Robert L. Axtel, *Science* 293, 1818–1820 (7 September 2001)

and means that idiosyncratic shocks of large firms may affect GDP (Gabaix 2006). Zipf's law holds for different measures of firm sizes and countries (Axtell 2001; Fujiwara et al. 2004; Gabaix and Landier 2008).

City sizes In the upper tail, Zipf's law holds generally well across times and countries (Gabaix and Ioannides 2004).

Gibrat's law for the growth rate of cities is shown in the United States by Ioannides and Overman (2003).

Roberts's law for executive compensation Across times and countries, an executive heading a firm of size S earns an amount proportional to S^{κ}, for a κ around 1=3. Superstars models explain the presence of this scaling (Gabaix and Landier 2006), but the reason for the 1/3 value remains a mystery.

More Recently Proposed Laws

Power law of stock market activity: returns, trading volume and trading frequency Following Mandelbrot, the following regularities have been found. Stock market returns (over one minute to one week) have PL tails, with an exponent around three (Gopikrishnan et al. 1999). Individual trades have a PL exponent around 1.5 (Gopikrishnan et al. 2000). The number of trades executed over a short horizon has an exponent close to three (Plerou et al. 2000). There is no consensus about the origins for those regularities. The fat tails of the returns might come from GARCH effects. One view (Gabaix et al. 2003, 2006) attributes it to the trades of large institutional investors in relatively illiquid markets, which creates spikes in returns and volume, and generates empirically found exponents.

Supply of regulations Mulligan and Shleifer (2005) establish another candidate law. In US states, the quantity of regulation is a PL of population.

Estimation of Power Laws

How does one estimate a distributional PL? We take the example of n cities in the upper tail, ordering them by size, $S_{(1)} \geq \cdots \geq S_{(n)}$. One method is Hill's estimator:

$$\hat{\zeta}^{Hill} = (n-1)/\sum_{i=1}^{n-1}\left(\ln S_{(i)} - \ln S_{(n)}\right)$$

which has a standard error $\hat{\zeta}^{Hill} n^{-1/2}$. The second method is a 'log rank log size regression', where $\hat{\zeta}$ is the slope in the regression of the log rank i on the log size:

$$\ln(i - s) = \text{constant} - \hat{\zeta}^{OLS} \ln S_{(i)} + \text{noise}$$

which has a standard error of $\hat{\zeta}^{OLS} \cdot (n/2)^{1/2} \cdot s$ is a shift, $s = 0$ is typical, but $s = 1/2$ is optimal (Gabaix and Ibragimov 2006). Both methods have pitfalls, as true errors are often larger than nominal standard errors (Embrechts et al. 1997; Gabaix and Ioannides 2004).

See Also

▶ ARCH Models
▶ Econophysics
▶ Inequality (Measurement)
▶ Quantity Theory of Money
▶ Superstars, Economics of
▶ Systems of Cities
▶ Wealth

Bibliography

Axtell, R. 2001. Zipf distribution of U.S. firm sizes. *Science* 293: 1818–1820.

Brown, J.H., and G.B. West. 2000. *Scaling in biology.* Oxford: Oxford University Press.

Champernowne, D. 1953. A model of income distribution. *Economic Journal* 83: 318–351.

Embrechts, P., C. Kluppelberg, and T. Mikosch. 1997. *Modelling extremal events for insurance and finance*. New York: Springer.

Fujiwara, Y., C. Di Guilmi, H. Aoyama, M. Gallegati, and W. Souma. 2004. Do Pareto–Zipf and Gibrat laws hold true? An analysis with European firms. *Physica A* 335: 197–216.

Gabaix, X. 1999. Zipf's law for cities: An explanation. *Quarterly Journal of Economics* 114: 739–767.

Gabaix, X. 2006. The granular origins of aggregate fluctuations. Working paper, MIT.

Gabaix, X., P. Gopikrishnan, V. Plerou, and H.E. Stanley. 2003. A theory of power law distributions in financial market fluctuations. *Nature* 423: 267–230.

Gabaix, X., P. Gopikrishnan, V. Plerou, and H.E. Stanley. 2006. Institutional investors and stock market volatility. *Quarterly Journal of Economics* 121: 461–504.

Gabaix, X., and R. Ibragimov. 2006. *Log (Rank-1/2): A simple way to improve the OLS estimation of tail exponents*. Working paper: Harvard University.

Gabaix, X., and Y. Ioannides. 2004. The evolution of the city size distributions. In *Handbook of regional and urban economics*, ed. V. Henderson and J.-F. Thisse, Vol. 4. Amsterdam: North-Holland.

Gabaix, X., and A. Landier. 2008. Why has CEO pay increased so much? *Quarterly Journal of Economics* 123: 49–100.

Gopikrishnan, P., V. Plerou, L. Amaral, M. Meyer, and H.E. Stanley. 1999. Scaling of the distribution of fluctuations of financial market indices. *Physical Review E* 60: 5305–5316.

Gopikrishnan, P., V. Plerou, X. Gabaix, and H.E. Stanley. 2000. Statistical properties of share volume traded in financial markets. *Physical Review E* 62: R4493–R4496.

Ioannides, Y.M., and H.G. Overman. 2003. Zipf's law for cities: An empirical examination. *Regional Science and Urban Economics* 33(2): 127–137.

Jessen, A.H., and T. Mikosch. 2006. Regularly varying functions. *Publications de l'Institut Mathématique, Nouvelle Série* 80: 171–192.

Luttmer, E.G.J. 2007. Selection, growth, and the size distribution of firms. *Quarterly Journal of Economics* 122(3).

Mulligan, C., and A. Shleifer. 2005. The extent of the market and the supply of regulation. *Quarterly Journal of Economics* 120: 1445–1473.

Plerou, V., P. Gopikrishnan, L. Amaral, X. Gabaix, and H.E. Stanley. 2000. Economic fluctuations and anomalous diffusion. *Physical Review E* 62(3): R3023–R3026.

Rossi-Hansberg, E., and M. Wright. 2007. Urban structure and growth. *Review of Economic Studies* 74: 597–624.

Schumpeter, J. 1949. Vilfredo Pareto (1848–1923). *Quarterly Journal of Economics* 63: 147–172.

Simon, H. 1955. On a class of skew distribution functions. *Biometrika* 44: 425–440.

Sornette, D. 2001. *Critical phenomena in natural sciences*. New York: Springer.

Power, Eileen Edna (1889–1940)

Phyllis Deane

Eileen Power's undergraduate years were spent at Girton College, Cambridge, where she was a teaching fellow in history for some years. Having little patience, however, with the medieval treatment Cambridge University meted out to female scholars, she was glad to escape first to the Sorbonne, then to the London School of Economics as a research student (1911–13) and finally on a travelling fellowship to the Far East in 1920–21. In 1921 she returned to the LSE, where she was rapidly promoted to Reader and in 1931 to Professor. There at least she did not have to fight for the privilege of giving her immensely popular lectures in a university lecture room.

The most distinctive feature of Eileen Power's contributions to economic history was her ability to distil from a pyramid of dusty manuscripts an account of human and institutional behaviour which was at once coherent, clear, comprehensive and lively. Her *Medieval English Nunneries* (1922) and *Medieval People* (1924) attracted a wide audience. Her posthumously published *Wool Trade in English Medieval History* (1941) synthesized two decades of patient detective work in contemporary records and literary sources and told the story of the medieval woollen industry from the pasture to the loom, the counting house and the ports. The detailed statistical research into the voluminous English trade records which she stimulated the members of her LSE seminars to exploit led to the pathbreaking *Studies in English Trade in the Fifteenth Century* (1933) which she edited with the young Michael Postan, her closest collaborator from 1926, when he became her research assistant, and her husband from 1937.

But it was through the stimulus of her imaginative scholarship on other economic historians that Eileen Power made her strongest impact on the study of economic history. It was said of her when she was being presented for an honorary degree at Manchester that she combined the graces of a butterfly with the sober industry of the bee. She was one of the founder members of the Economic History Society and the wide range of her international professional friendships made her an invaluable member of the Editorial Board of the *Economic History Review.*

See Also

▶ Postan, Michael Moïssey (1899–1981)

Selected Works

1920. *The Paycockes of Coggeshall*. London: Methuen.

1922. *Medieval English nunneries*. Cambridge: Cambridge University Press.

1924. *Medieval people*. London: Methuen.

1933. (Ed., with M.M. Postan.) *Studies in English trade in the fifteenth century.* London: Routledge.

1941. *The wool trade in English medieval history.* Oxford: Clarendon Press.

Pownall, Thomas (1722–1805)

Henry W. Spiegel

English colonial administrator and Governor of Massachusetts, Pownall was an early critic of Adam Smith's *Wealth of Nations.* He published his criticism under the title *A Letter from Governor Pownall to Adam Smith* in 48 large quarto pages late in 1776. Smith acknowledged it in a polite letter early in 1777. There is a further reference to the matter in Smith's correspondence with Andreas Holt of 26 October 1780.

In the *Letter,* Pownall expresses his admiration for Smith's work as a whole, which, if properly corrected, should serve as the basis of lectures at the universities. He finds fault, however, with a number of Smith's ideas. He opposes Smith's view of the propensity to barter as a cause of the division of labour. To him, the latter stems from differences in men's natural endowment. He has doubts about Smith's distinction between natural and market price, the latter being the only 'real' price, and about the search for a real measure of value. He does not accept the view that relative prices are measured by labour; in his opinion they reflect the bargaining power of the parties. He defends the monopoly of colonial trade, bounties on exports and restraints on imports, and has much praise for metallic money, which is a national asset and not merely a medium of exchange. He takes Smith to task for underestimating the economic and military value of the North American colonies and for proposing the dismemberment of the empire. Some of Pownall's critical points are informed by his experience as a colonial administrator. His criticism, on the whole, reflects a late flowering of mercantilist views and demonstrates that Smith's ideas did not altogether go unchallenged by exponents of the old order.

Pownall's *Letter* was only a passing episode in a life filled with accomplishments in public service, and he was the author of *The Administration of the Colonies,* first published in 1764 and frequently reprinted, in which he proposes a commonwealth-type colonial reorganization.

Selected Works

1776. *A letter from governor Pownall to Adam Smith.* Facsimile reprint. New York: Augustus Kelley, 1967. Also reprinted as Appendix A, in *The correspondence of Adam Smith,* ed. E.-C. Mossner and I.S. Ross. Oxford: Clarendon Press, 1977.

References

Pownall, C.A.W. 1908. *Thomas Pownall*. London: Stevens.
Schutz, J.A. 1951. *Thomas Pownall*. Glendale: A.H. Clark.

Prebisch, Raúl (1901–1986)

José Gabriel Palma

Abstract

Prebisch was concerned with four stylized facts of commodity-exporting middle- income countries: (*a*) failure to 'catch-up'; (*b*) recurrent balance of payments disequilibrium; (*c*) unstable and deteriorating terms of trade; and (*d*) persistent unemployment. At the core of Prebisch's analysis lies his differentiation of the economic and export structures of the centre and periphery. Those of the centre were seen as *homogeneous* and *diversified,* those of the periphery as *heterogeneous* and *over-specialized.* Prebisch associates the second and third problems above primarily with export over-specialization (too few homogeneous, unbranded and price-volatile commodities), and the first and fourth with structural heterogeneity (which hindered industrialization).

Keywords

Balance of payments; Commodity-exporting countries; Comparative advantage; Cooperative games; Excess demand; Kaldor, N.; Import substitution; Industrialization; Long-term economic growth; Periphery; Prebisch, R.; Robinson, J. V.; Structuralism; Terms of trade

JEL Classifications

B31

Prebisch was born on 17 April 1901 in Tucumán, Argentina, and died at the age of 84 in Santiago de Chile. He graduated in Economics at the University of Buenos Aires in 1923 having already published six papers in academic journals.

He was Professor of Political Economy at the University of Buenos Aires from 1925 to 1948. In 1930 he became Under-Secretary of Finance at the age of 29, and soon afterwards the first Director General of the Argentine Central Bank (1935–43). He then moved to the United Nations, being appointed Executive Secretary of the Economic Commission for Latin America and the Caribbean (ECLAC) in 1950. In 1963 he moved to the United Nations Conference on Trade and Development (UNCTAD) as its first Secretary General.

Although his main intellectual concern was always the understanding of the specific development obstacles facing commodity-exporting, middle-income, peripheral countries, he always acknowledged that early in his career he had viewed them from a mainstream perspective. His approach only changed when he witnessed the Great Depression (including the heterodox response to it of many industrialized countries) and read the *General Theory.* After writing several articles and an influential book on Keynes, in the 1950s he led his ECLAC team (which eventually included Fernando Henrique Cardoso, Enso

Faletto, Celso Furtado, Aníbal Pinto and Osvaldo Sunkel) in the formulation of the 'structuralist approach' (see dependency; Furtado, Celso; structuralism).

In this approach, Prebisch was basically concerned with four stylized facts of commodity-exporting, middle-income countries: (*a*) their growing income gap with industrialized countries (failure to 'catch up'); (*b*) their recurrent balance of payments disequilibrium; (*c*) the instability and the tendency to deterioration of their terms of trade; and (*d*) their persistent unemployment (often coexisting with inflationary pressures). At the core of Prebisch's analysis lies his differentiation of the economic and export structures of the centre and the periphery. Those of the centre were seen as *homogeneous* and *diversified,* those of the periphery as *heterogeneous* and *over-specialized*. Heterogeneous because economic activities with remarkably different productivity-growth dynamics existed side by side – namely, a modern export sector coexisting with a backward agriculture and an undersized manufacturing sector. Over-specialized because the range of exports was limited to just a few (homogenous, unbranded and price-volatile) commodities, and their process of production had very limited backward- and forward-linkages with the rest of the economy (see structuralism).

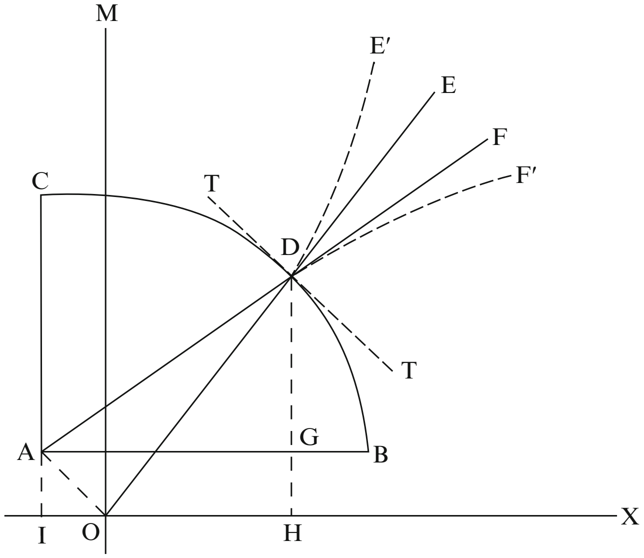
The recurrent cyclical problems of the balance of payments and the instability and the tendency to deterioration of the terms of trade are associated primarily with an excessive degree of export specialization (due to a narrow Ricardian understanding of comparative advantage); the problems of slow growth (and failure to 'catch up' with industrialized countries) and persistent unemployment with the constraints created by structural heterogeneity interacting with export overspecialization (which, among other things, hindered industrialization and created inflationary pressures).

His best-known thesis is the tendency to deterioration in the terms of trade of the periphery, the development of which coincided with (and owed much to) Hans Singer's work (1950). It is not clear whether he saw this as his most important contribution (or even as the most significant problem of a commodity-exporting country), but by its own nature the 'Prebisch–Singer' thesis was a seductive empirical challenge to that part of the academic world which is ever anxious for onedimensional hypotheses referring to clearly established variables. Prebisch was in fact much more concerned with the lack of impetus for industrialization resulting from a narrow and static Ricardian integration into the world economy. His main hypothesis is that there are reinforcing elements that – left to unregulated markets – would tend to work against the periphery's growth and welfare.

The tendency towards deterioration in the terms of trade of the periphery could be synthesized as in Fig. 1.

From the point of view of the periphery's consumption path, the Prebisch-Singer hypothesis is that the income elasticity of the periphery's imports from the centre (manufactures) is not only greater than 1 but also much greater than that of the centre for products of the periphery (commodities). Therefore, left to unregulated markets the long-term trend of the periphery's consumption path would be biased towards trade with the centre (say, ODE′ instead of the 'trade-neutral' ODE); that is, as incomes grow the proportion of importables (from the centre) in the periphery's consumption would increase. This would not be the case for the centre in terms of the share of its commodity imports from the periphery in total consumption (their income and price elasticities for commodities are low).

The same long-term trade bias would tend to happen in the production path – the periphery would tend to move along the ADF′ path instead of the 'trade-neutral' ADF one. That is, as output grows the share of (low price- and low income-elasticity) commodities for export in total output would increase, not least because of the additional foreign exchange needed to finance the trade-biased consumption path (something that often turns out to be a self-defeating endeavour). Therefore, vis-à-vis each other's products, the periphery would tend to have a more trade-biased path than the centre in terms of both consumption and production. There would consequently be a tendency for an excess demand from the periphery for imported manufactures from the centre, and an



Prebisch, Raúl (1901–1986), Fig. 1 X = exportable of the periphery (Primary commodity); M = importable of the periphery (manufacturing good); ABC = transformation curve of the periphery; ODE = the periphery's 'neutral' consumption path; ODE' = its 'biased-for-trade' consumption path; ADF = the periphery's 'neutral' production path; ADF' = its 'biased-for-trade' production path. For the periphery, at point D, OA = TT = terms of trade, DH = consumption of M, DG = local production of M, GH = AI = imports of M, IH = production of X, OH = local consumption of X; and IO = exports of X

excess supply of commodities, resulting in a tendency towards a deterioration of the terms of trade.

This tendency would be reinforced because of a similarity and a difference in productivity growth between commodities and manufactures. The similarity is that productivity growth can be relatively high in both types of products (commodities and manufactures, although is likely to be faster in the latter). The difference is that productivity increases in the centre's manufacturing do not tend to be transferred into lower prices as much as those of (homogenous and unbranded) primary production in the periphery (due to market imperfections in the centre's product and labour markets, mainly oligopolistic firms operating in product-differentiated markets and strong unions).

The end result would be that, if both poles were to grow at the same rate of per capita income, the periphery's more trade-biased path in both consumption and production (vis-à-vis each other's products) would tend to generate a deficit in its trade balance with the centre. Therefore, a long-term equilibrium in their balance of payments would require the income per capita of the periphery to grow systematically at a lower rate than that of the centre (the opposite of a 'catching up' scenario).

Further, Prebisch adds that this growth asymmetry would be reinforced by the fact that productivity growth in manufactures tends to have higher positive externalities and spillover effects, stronger linkages with the domestic economy, steeper technology ladders, and so on. Within this context, Prebisch argues that (for reasons of supply and demand) the periphery could achieve a higher sustainable growth path only by substituting highly income-elastic manufacturing imports with domestic production, and diversifying its exports towards more income- and price- elastic, productivity-enhancing products – that is, if it were to embark on a deeper and faster process of industrialization than one that would 'spontaneously' emerge from a Ricardian integration into the world economy.

Therefore, Prebisch's arguments for forcing the pace of industrialization are not only based on differences in income elasticities of demand for imports and in price elasticities of demand for exports (arguments at the level of the circulation of commodities), but are also due to the growth-enhancing nature of manufacturing activities (an argument at the level of production; see Kaldor, Nicholas).

Prebisch’s theory challenges Ricardo’s comparative advantage premises – in fact, for Prebisch the higher the rate of growth of productivity in the periphery’s primary commodity export sector, the greater the need for import-substituting industrialization (1983, p. 1082). It also challenges the classical terms-of-trade approach – for example, Mill (1848) and Keynes (1920) – which argued that in the long term they are bound to move in favour of commodities (mainly due to hypothetical diminishing returns in commodity production). Prebisch’s logic would later influence Joan Robinson (1979) when she argued that in Ricardo’s example Portugal ends up with a low rate of accumulation, and having destroyed its promising textile industry, while England in contrast had an industrial revolution (see Robinson, Joan Violet). It would also influence Kaldor’s arguments in favour of manufacturing-led growth (1967), Pasinetti’s multi-sector macro-dynamics framework (1983), Ajit Singh’s concept of an optimal degree of industrialization (1977), Rowthorn and Wells’ seminal work on de-industrialization (1987), and Thirlwall’s balance-of-payments-constrained growth multiplier (2003).

For criticisms of Prebisch’s ideas, see structuralism and dependency. Some additional issues to which the literature on Prebisch has not given due consideration are as follows:

1. Although there is no evidence that this was Prebisch’s intention, for many years his ideas led in many intellectual and policymaking circles to a strong bias against commodity production per se.
2. The asymmetric trade liberalization that has taken place since globalization (the periphery opening up to manufacturing imports, but the centre not reciprocating for commodities) has deepened the problems identified by Prebisch.
3. Prebisch’s Argentinian background is undoubtedly responsible for his focusing mainly on the relative decline of middle-income, commodity-rich countries, and for his scepticism regarding the role of an inelastic supply of agricultural products in explaining the regions’ persistent inflationary pressures (Argentina simply did not fit the pattern of the structuralist theories of inflation developed by many of his colleagues at ECLAC; see structuralism).
4. The recent remarkable export drive of basic (homogenous and unbranded) manufacturing products in some developing countries has led to their ‘price commoditization’ , leading to a similar terms of trade problem vis-à-vis their imports of more technologically advanced manufactures (see Palma 2005).
5. There are significant fallacy-of-composition issues among commodity-exporting countries (for example, actual price elasticity of demand crucially depends on market shares), making cooperative games among producers difficult.
6. Probably because of his own ‘institutional constraints’ (inevitable when working in international organizations), Prebisch never addressed properly some crucial institutional issues associated with the often poor macroeconomic performance of many mineral-exporting economies – such as those analysed (not always successfully) by the ‘resource course’ literature (see de-industrialization, ‘premature’ de-industrialization and the Dutch Disease).
7. Prebisch’s preferences for a ‘stage’ approach (first an import-substituting phase, then a manufacturing export-oriented one towards a regional custom union, then one to the rest of the world) did not account for institutional and political path- dependency inertias that would create almost insurmountable hurdles for the transition even to the second stage; East Asia’s ‘simultaneous’ approach was far more successful (Palma 2007).
8. Finally (and crucially), the contrasting experiences of Latin America and East Asia show that, while it is one thing to use trade and industrial policies to create incentives (rents)

to divert resources towards more 'dynamic' products (that is, income- and price-elastic manufacturing products, with deeper linkages, higher productivity growth potential, stronger externalities and spillover effects, useful for technology ladders, and so on), it is quite another to have the institutional capabilities necessary to ensure that the capitalist elite uses those rents effectively (Khan 2000).

See Also

De-Industrialization, 'Premature' De-industrialization and the Dutch Disease
- ▶ Dependency
- ▶ Díaz-Alejandro, Carlos (1937–1985)
- ▶ Furtado, Celso (1920–2004)
- ▶ Kaldor, Nicholas (1908–1986)
- ▶ Robinson, Joan Violet (1903–1983)
- ▶ Structuralism
- ▶ Terms of Trade

Selected Works

1949. The economic development of Latin America and its principal problems. *Economic Bulletin for Latin America* 7, 1962, 1–22 (first published by ECLAC in 1949).

1951a. The spread of technical progress and the terms of trade. In *Economic Survey of Latin America, 1949*. New York: UN-DESA.

1951b. *Problemas teóricos y prácticos del crecimiento económico*. Santiago: UN-ECLAC.

1959. Commercial policy in the underdeveloped countries. *American Economic Review* 49: 251–273.

1962. El falso dilema entre desarrollo económico y estabilidad monetaria. *Boletín Económico de América Latina* 6(1): 1–26.

1963. *Towards a dynamic development policy for Latin America*. New York: UN.

1968. A new strategy for development. *Journal of Economic Studies* 3(1): 3–14.

1971a. *Latin America: A problem in development*. Austin: Institute of Latin American Studies, University of Texas.

1971b. *Change and development – Latin America's great task*. New York: Praeger; Mexico: Fondo de Cultura Económico.

1976. A critique of peripheral capitalism. *CEPAL Review* 1: 9–76.

1981. Capitalismo periférico: crisis y transformación. *México: Fondo de Cultura Económico.*

1983. Tres etapas de mi pensamiento económico. *El Trimestre Económico* 50: 1077–1096.

Bibliography

Di Marco, L.E. 1972. The evolution of Prebisch's economic thought. In *International economics and development: Essays in honour of Raul Prebisch*, ed. L.E. Di Marco. New York: Academic.

Kaldor, N. 1967. Problems of industrialization in underdeveloped countries. In *Strategic factors of economic development*. Cornell: Cornell University Press.

Keynes, J.M. 1920. *Economic consequences of the peace*. London: Macmillan.

Khan, M. 2000. Rent-seeking as process. In *Rents, rent-seeking and economic development*, ed. M. Khan and K. Jomo. Cambridge: Cambridge University Press.

Love, J. 1995. Economic ideas and ideologies in Latin America since 1930. In *Cambridge history of Latin America*. vol. 6, part I. Cambridge: Cambridge University Press.

Mill, J.S. 1848. *Principles of political economy*. London: Longmans.

Ocampo, J.A., and M.A. Parra. 2003. The terms of trade for commodities in the twentieth century. *ECLAC Review* 29: 7–35.

Palma, J.G. 2005. The six main stylised facts of the Mexican economy since NAFTA. *Journal of Industrial and Corporate Change* 14: 941–991.

Palma, J.G. 2007. Flying geese and waddling ducks: The different capabilities of East Asia and Latin America to 'demand-adapt' and 'supply upgrade' their export productive capacity. In *Industrial policy in developing countries*, ed. M. Cimoli, G. Dosi, and J. Stiglitz. Oxford: Oxford University Press.

Pasinetti, L. 1983. The accumulation of capital. *Cambridge Journal of Economics* 7: 405–411.

Robinson, J.V. 1979. *Aspects of development and underdevelopment*. Cambridge: Cambridge University Press.

Rodriguez, O. 1980. *La teoría del subdesarrollo de la CEPAL*. Mexico: Siglo XXI Editores.

Rodriguez, O. 2007. *El estructuralismo latinoamericano*. Mexico: Siglo XXI Editores.

Rowthorn, R., and J. Wells. 1987. *De-industrialisation and foreign trade*. Cambridge: Cambridge University Press.

Singer, H. 1950. The distribution of gains between investing and borrowing countries. *American Economic Review* 40: 473–485.

Singh, A. 1977. UK industry and the world economy: a case of de-industrialisation? *Cambridge Journal of Economics* 1: 113–116.

Thirlwall, A. 2003. *Trade, the balance of payments and exchange rate policy in developing countries*. Cheltenham: Edward Elgar.

Precautionary Principle

Christian Gollier and Nicolas Treich

Abstract

The precautionary principle (PP), as it appears in international treaties or in some countries' legal systems, suggests that the prospect of scientific progress should not justify the delay of preventive measures. Three effects identified in the economics literature – the irreversibility, the precautionary and the ambiguity aversion effects – may be consistent with the normative content of the PP. A difficult question is how then the PP can be implemented. Several social actors may want to take advantage of a current lack of scientific evidence to promote their own interests. The PP can also be misused, for example, for demagogy or protectionism.

Keywords

Ambiguity and ambiguity aversion; Asymmetric information; Availability heuristic; Cartagena Protocol on Biosafety (2000); Convention on Biological Diversity (1992); Convention on Climate Change (1992); Environmental economics; Irreversibility effect; Maastricht Treaty (EU); Precautionary effect; Precautionary principle; Protection; Risk; Risk perception; Scientific progress; Scientific uncertainty

JEL Classifications

D81; Q50; H43; C44

The precautionary principle (PP) is a recent notion. It has its roots, some believe, in the early 1970s as the German principle of *Vorsorge,* or foresight (see, for example, O'Riordan and Cameron 1994; Morris 2000; Sunstein 2005). It is often said that the PP was first introduced in 1984 at the International Conference on Protection of the North Sea. Its popularity increased after the Conference of Rio in 1992; Principle 15 of the Rio Declaration states, 'Where there are threats of serious or irreversible damage, lack of full scientific certainty shall not be used as a reason for postponing cost-effective measures to prevent environmental degradation' (UNGA 1992).

Similar definitions have been proposed in international statements of policy, including the 1992 Convention on Climate Change, the 1992 Convention on Biological Diversity, the Maastricht Treaty in 1992/1993, and the 2000 Cartagena Protocol on Biosafety. The PP has also been enacted in the national law of several countries, especially in Europe.

The PP is the most notable anticipatory principle with special relevance for human-induced environmental problems under conditions of scientific uncertainty. Although devoid of practical content, the main message of the PP is conceptually clear: the prospect of anticipation of scientific progress should not justify the delay of measures preventing environmental degradation. In practice, its scope has became wider and there are reasonable grounds for applying it to regulate the protection of human, animal and plant health issues (see, for example, Commission of European Communities 2000).

The economic analysis of the PP has mostly studied the tension between two effects: (*a*) developing an economic activity that is profitable now but may pose risks to the society in the future, and (*b*) not developing this activity until conclusive scientific information is forthcoming about its harmlessness. The PP is said to be socially efficient if the benefit of postponing the risky activity is greater than its cost. To put it differently, the PP is efficient if the net social benefit of early prevention efforts is positive. The economic conditions for efficiency were first analysed in the 1970s in the literature on the 'irreversibility effect'.

The Irreversibility Effect

Let us consider a model of economic decisions represented by the following optimization program

$$\max_{x_1 \in D_1} E_{\tilde{y}} \max_{x_2 \in D(x_1)} E_{\tilde{\theta}/\tilde{y}}\ v\left(x_1, x_2, \tilde{\theta}\right) \tag{1}$$

The timing of the model is the following. At date 1, the decision-maker chooses x_1 in a set D_1. Between date 1 and date 2, he observes the realization of a signal y correlated to $\tilde{\theta}$. At date 2, before the realization of $\tilde{\theta}$, he chooses x_2 in a set $D(x_1)$. Finally $\tilde{\theta}$ is realized and the decision-maker gets a utility payoff $v(x_1; x_2, \theta)$. The problem is to determine the effect of a ‘better information structure’ $\tilde{y}$ on the optimal decision at date 1.

We first solve this problem when $v(x_1, x_2, \theta) = x_1 + x_2\theta$ with $D_1 = \{0, 1\}$ and $D(x_1) = \{x_1, 1\}$. This special case can be interpreted as a simple investment problem. The development of a ‘risky’ project – like the exploitation of a forest in which the value of biodiversity is unknown – is considered. If the project is implemented today ($x_1 = 1$), it yields a net benefit of 1 today and of $\tilde{\theta}$ in the future. The project is irreversible in the sense that once it is developed it cannot be stopped ($x_2 = 1$ if $x_1 = 1$). The stakeholders are assumed to be risk neutral.

Consider first the case in which no scientific progress is expected, that is, when $\tilde{y}$ is independent of $\tilde{\theta}$. In this case, program (1) becomes

$$\begin{aligned}&\max_{x_1 \in \{0,1\}, x_2 \in \{x_1, 1\}} E_{\tilde{\theta}}\left(x_1 + x_2\tilde{\theta}\right)\\ &= \max\left(1 + E_{\tilde{\theta}}\tilde{\theta}, 0\right)\end{aligned} \tag{2}$$

Either the project is implemented today if its expected net present value (ENPV) is positive, that is if $1 + E_{\tilde{\theta}}\tilde{\theta} \geq 0$, or it is never implemented. Consider alternatively the case of scientific progress that yields perfect information about $\tilde{\theta}$. This is equivalent to assuming perfect correlation between $\tilde{y}$ and $\tilde{\theta}$. In this case, program (1) becomes

$$\begin{aligned}&\max_{x_1 \in \{0,1\}} E_{\tilde{\theta}} \max_{x_2 \in \{x_1, 1\}}\left(x_1 + x_2\tilde{\theta}\right)\\ &= \max\left(1 + E_{\tilde{\theta}}\tilde{\theta}, E_{\tilde{\theta}}\max\left(0, \tilde{\theta}\right)\right)\end{aligned} \tag{3}$$

Viewed today, the ENPV of postponing the decision to develop the project equals $V = E_{\tilde{\theta}}\max\left(0, \tilde{\theta}\right)$. The project will be initiated today only if it yields a larger ENPV than that obtained if the decision is postponed to the future: $1 + E_{\tilde{\theta}}\tilde{\theta} \geq V$. The quantity V has been coined the (quasi-)option value (Arrow and Fischer 1974).

The comparison between (2) and (3) shows that scientific progress has the effect of increasing the ENPV of the best alternative option from 0 to $V \geq 0$. Consistent with the PP, this example shows that the prospect of scientific progress may lead to the postponement of the development of the risky project. The prospect of receiving information in the future increases the cost of choosing the irreversible decision today. This decision would prevent the decision maker from taking advantage of the information in the future. This is the ‘irreversibility effect’ (Henry 1974).

The literature has studied the generalization of this effect in several directions, including partial resolution of uncertainty, relative flexibility, continuous decision variables, non-separable preferences and risk aversion. This example relied on two extreme information structures: one structure gives no information and the other gives perfect information. The appropriate general notion of a ‘better information structure’ was introduced by Blackwell (1951). This general notion was used and developed in a systematic way by Epstein (1980) under some differentiability assumptions. Epstein then demonstrated that the irreversibility effect does not hold for most payoff functions $v(x_1; x_2, \theta)$. Jones and Ostroy (1984) have generalized Epstein’s result to non-differentiable problems and to a more general characterization of adjustment costs.

The Precautionary Effect

The subsequent literature has mostly used Epstein’s approach to examine the effect of better information for various payoff functions, on the assumption of continuous decisions, differentiability and that the conditions for optimization in (1) were satisfied. Ulph and Ulph (1997) consider a payoff function of the form $v(x_{1,} x_2, \theta) = u_1 (x_1) + u_2(x_2) - \theta d(\delta x_1 + x_2)$ and interpret x_t as the emissions of CO_2 in period t and $\theta d(.)$ as the uncertain climate damage that depends on the sum of emissions up to a decay parameter δ. They show that a better information structure may lead to an increase, not a decrease, in emissions at date 1. Gollier et al. (2000) analyse a similar model with monetary damages $v(x_1, x_2, \theta) = u_1 (x_1) + u_2(x_2 - \theta(\delta x_1 + x_2))$. They show that that emissions at date 1 decrease if and only if $u_2(.)$ has a constant relative risk aversion lower than 1, or a derivative 'sufficiently' convex. This latter condition suggests that the coefficient of prudence (Kimball 1990) is instrumental in signing the effect of a better information structure on x_1. This is not surprising since in this model x_1 affects future utility u_2, no longer by reducing the future set of choices but directly by changing the risk borne in the future $\theta(\delta x_1 + x_2)$. This is the 'precautionary effect'. Overall these results suggest that the qualitative effect of a better information structure strongly depends on functional forms, in particular on the risk attitude of the decision maker.

The Ambiguity Aversion Effect

The Ellsberg paradox tells us that many people do not behave according to the expected utility criterion when facing (scientific) uncertainty, contrary to what we assumed above. Gilboa and Schmeidler (1989) proposed an alternative decision criterion that performs better in this context. Under their model of ambiguity aversion, for each possible choice *ex ante,* the decision maker computes the expected utility conditional to each plausible scientific theory, and takes the minimum to evaluate the welfare generated by that choice. Agents who behave according to this maxmin model exhibit a form of choice-sensitive pessimism, which is called 'ambiguity aversion'. As shown for example by Chen and Epstein (2002) for financial markets, this ambiguity aversion reinforces risk aversion to induce people to adopt a more precautionary behaviour in the case of (scientific) uncertainty, as suggested by the PP.

Positive Aspects of the Precautionary Principle

The economic approach of the PP has been mostly normative so far. Under which conditions is the PP socially efficient? How should scientific uncertainty affect risk management? An equally important approach involves discussion on how the PP has been or should be implemented. We briefly turn to these more positive aspects.

A general argument is that scientific uncertainty may exacerbate, or even trigger, some market or regulatory failures (Gollier and Treich 2003). With a global pollution problem such as climate change, there are incentives for countries to free ride on other countries' reduction of emissions. Coalitions formations may reduce this inefficiency but coalitions are less likely to form if there is scientific progress (Na and Shin 1998). At a political level, an argument used by governments is that the problem is 'too uncertain' to abate emissions. Early commitments may help, but there are incentives for some governments, once information reveals low levels of damage in their own country, to refuse to abate emissions at a level announced by previous governments.

A difficult question is that of the most efficient policy to induce firms to internalize the risks they pose for the economy. In a market with imperfect legally enforceable property rights, firms may not take up the option of waiting for better information when high profits are guaranteed to first-movers. How to set binding legal incentives for firms' past actions made under conditions of scientific uncertainty is a big issue in law. This issue is augmented by the classical limited liability problem.

Another issue is that of international relations and the different approaches to safety and precaution across countries (Hammitt et al. 2005). One possibility is to leave states to decide how to account for scientific uncertainty in their safety policy. The problem is that such a discretionary power may be the source of disguised protectionism.

Scientific uncertainty may also increase the cognitive biases of the public in their perception of risks, like the standard 'availability heuristic'. Citizens often deem an event to be more probable when its occurrence can be easily recalled or visualized. As a result, they may overreact to highly publicized risks. Interest groups may exploit this bias, as well as politicians. A critical interpretation of the PP is to view it as a demagogic response to citizens' perceptions of risks (Sunstein 2005).

More generally, scientific uncertainty may favour, through the multiple channels of decision-making, opportunistic behaviours. Scientific uncertainty creates space for discretion in the risk regulatory process. Several social actors (entrepreneurs, lobbies, experts, politicians, media, and so on) may take advantage of the lack of scientific evidence to promote their own interests. The PP may be viewed as a soft safeguard against opportunistic behaviours in situations of asymmetric and evolving information. Yet designing stronger mechanisms needs a more detailed analysis of the sources of market failures, of risk management institutions and of citizens' behavioural responses. This may partly explain the existing voluminous literature on the PP in the social sciences, and may occupy economists in the future.

See Also

- ▶ Ambiguity and Ambiguity Aversion
- ▶ Cost–Benefit Analysis
- ▶ Irreversible investment
- ▶ Risk

Bibliography

Arrow, K., and A. Fischer. 1974. Environmental preservation, uncertainty and irreversibility. *Quarterly Journal of Economics* 88: 312–319.

Blackwell, D. 1951. Comparison of experiments. In *Proceedings of the second Berkeley symposium on mathematical statistics and probability*, ed. J. Neyman. Berkeley: University of California Press.

Chen, Z., and L. Epstein. 2002. Ambiguity, risk, and asset returns in continuous time. *Econometrica* 70: 1403–1443.

Commission of the European Communities. 2000. *Communication from the Commission on the Precautionary Principle.* Brussels: Commission of the European Communities. Online. Available at http://europa.eu.int/eur-lex/en/com/cnc/2000/ com2000_0001en01.pdf. Accessed 23 Mar 2006.

Epstein, L. 1980. Decision-making and the temporal resolution of uncertainty. *International Economic Review* 21: 269–284.

Gilboa, I., and D. Schmeidler. 1989. Maximin expected utility with non-unique prior. *Journal of Mathematical Economics* 18: 141–153.

Gollier, C., B. Jullien, and N. Treich. 2000. Scientific progress and irreversibility: An economic interpretation of the Precautionary Principle. *Journal of Public Economics* 75: 229–253.

Gollier, C., and N. Treich. 2003. Decision-making under scientific uncertainty: The economics of the Precautionary Principle. *Journal of Risk and Uncertainty* 27: 77–103.

Hammitt, J., J. Wiener, B. Swedlow, D. Kall, and Z. Zhou. 2005. Precautionary regulation in Europe and in the United States: A quantitative comparison. *Risk Analysis* 25: 1215–1228.

Henry, C. 1974. Investment decisions under uncertainty: The 'irreversibility effect'. *American Economic Review* 64: 1006–1012.

Jones, J., and R. Ostroy. 1984. Flexibility and uncertainty. *Review of Economic Studies* 6: 13–32.

Kimball, M. 1990. Precautionary savings in the small and in the large. *Econometrica* 61: 53–73.

Morris, J. 2000. Defining the precautionary principle. In *Rethinking risk and the precautionary principle*, ed. J. Morris. Oxford: Butterworth-Heinemann.

Na, S., and H. Shin. 1998. International environmental agreements under uncertainty. *Oxford Economic Papers* 50: 173–185.

O'Riordan, T., and J. Cameron, eds. 1994. *Interpreting the precautionary principle*. London: Earthscan Publications.

Sunstein, C. 2005. *Laws of fear: Beyond the precautionary principle*. Cambridge: Cambridge University Press.

Ulph, A., and D. Ulph. 1997. Global warming, irreversibility and learning. *Economic Journal* 107: 636–650.

UNGA (United Nations General Assembly). 1992. *Report of the United Nations conference on environment and development. Annex I: Rio Declaration on Environment and Development.* Rio de Janeiro, 3–14 June. A/CONF.151/26 (Vol. I) of 12 August. Online. Available at http://www.un.org/documents/ga/conf151/aconf15126-1annex1.htm. Accessed 30 Mar 2006.

Precautionary Saving and Precautionary Wealth

Christopher D. Carroll and Miles S. Kimball

Abstract

Precautionary saving measures the consequences of uncertainty for the rate of change (and therefore the level) of wealth. The qualitative aspects of precautionary saving theory are now well established: an increase in uncertainty will increase the level of saving, but will reduce the marginal propensity to save. Quantitatively, theory combined with empirical estimates of risk aversion suggests that precautionary saving and precautionary wealth should be quite large. More direct empirical evidence on precautionary saving suggests that precautionary effects on saving are substantial, but the magnitude of the effects is disputed, and the different estimates are not all expressed in comparable units.

Keywords

Calibration; Consumption function; Elasticity of intertemporal substitution; Euler equations; Impatience; Liquidity constraints; Perfect foresight; Precautionary saving; Precautionary savings; Precautionary wealth; Preferences; Risk aversion; Time consistency; Uncertainty

JEL Classifications

D4; D10

Precautionary saving is additional saving that results from the knowledge that the future is uncertain.

In principle, additional saving can be achieved either by consuming less or by working more; here, we follow most of the literature in neglecting the 'working more' channel by treating non-capital income as exogenous.

Before proceeding, a terminological clarification is in order. 'Precautionary saving' and 'precautionary savings' are often (understandably) confused. 'Precautionary saving' is a response of current spending to future risk, conditional on current circumstances. 'Precautionary savings' is the additional wealth owned at a given point in time as the result of past precautionary behaviour. That is, precautionary savings at any date is the stock of extra wealth that has resulted from the past flow of precautionary saving. To avoid confusion, we advocate use of the phrase 'precautionary wealth' in place of 'precautionary savings'.

Strength of the Precautionary Saving Motive

In the standard analysis, originally formulated in a two-period model by Leland (1968), and extended to the multi-period case by Sibley (1975) and Miller (1976), precautionary saving is modelled as the outcome of a consumer's optimizing choice of how to allocate existing resources between the present and the future. Additional interest in precautionary saving was stimulated by numerical solution of a benchmark model by Zeldes (1989) and the connection made in Barsky et al. (1986) between precautionary saving and the effects of government debt. (We assume time-invariant preferences in order to sidestep the important issues of time consistency recently explored by Laibson 1997, and others. That literature opens up a rich and interesting field of further behavioural possibilities beyond the basic logic outlined here.)

To clarify the theoretical issues, we break down the consumer's problem into two steps: the transition between periods, and the choice within the period. A consumer who ends period t with assets a_t receives capital income in period $t + 1$ of $a_t r$. The consumer's immediate resources ('cash-on-hand') in period $t + 1$ consist of such capital income, plus the assets that generated it, plus labour income y_{t+1}:

$$m_{t+1} = a_t r + a_t + y_{t+1} \tag{1}$$

$$= \underbrace{(1+r)}_{\equiv R} a_t + y_t + 1. \tag{2}$$

The simplest interpretation of m is as the contents of the consumer's bank account immediately after receipt of the paycheck and interest income ('cash-on-hand'). R is the real interest *factor*, as distinct from the real interest *rate*, lower case r. a_t reflects the consumer's accumulated assets at the end of period t, after the spending decision for period t has been made. The transition from the beginning to the end of period t reflects the fact that spending is paid for by drawing down m:

$$a_t = m_t - c_t. \tag{3}$$

To decide how to behave optimally in period t, the consumer must be able to judge the value of arriving in period $t + 1$ in any possible circumstance. This information is captured by the value function $v_{t+1}(m_{t+1})$. Here, we simply assume the existence of some well-behaved v_{t+1}; below we show how to construct v_{t+1}.

Standard practice assumes that consumers in period t weight future value by the factor β; if $\beta = 1$ the consumer today cares equally about current and future pleasure, while if $\beta < 1$ the consumer prefers present to future pleasure. Given β, and assuming that the consumer's period-t beliefs about future distribution of income are captured by the expectations operator E_t, we can define the value of ending period t with accumulated assets a_t as

$$\omega_t(a_t) = \beta E_t\left[v_{t+1}\left(Ra_t + \tilde{y}_{t+1}\right)\right], \tag{4}$$

where the ~ over the y indicates that period- $(t + 1)$ income is uncertain from the perspective of period t. Think of $\omega_t(a)$ as the end-of-period value function.

The consumer's goal is to optimally allocate beginning-of-period resources between current consumption and end-of-period assets; the value function for period t is defined as the function that yields the value associated with the optimal choice:

$$v_t(m_t) = \max_{c_t} \{u(c_t) + \omega_t(m_t - c_t)\}. \tag{5}$$

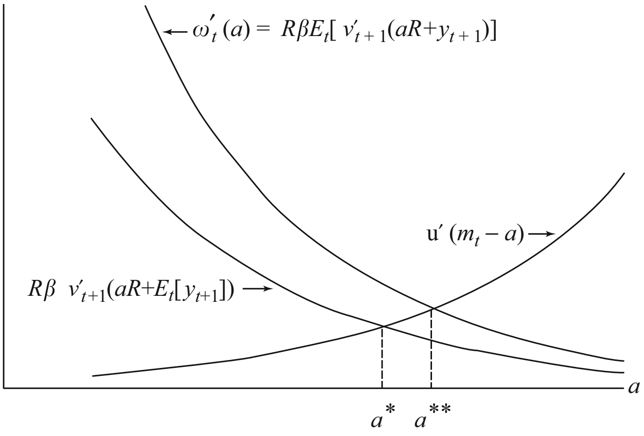
By definition, the optimal choice will be a level of c_t such that the consumer does not wish to change spending. Under standard assumptions, this implies that the marginal utility of consumption must be equal to the marginal value of assets:

$$u'\left(\overbrace{m_t - a_t}^{c_t}\right) = \omega_t'(a_t), \tag{6}$$

since if this were not true the consumer would be able to improve his well-being (value) by reallocating some resources between a and c.

Figure 1 depicts the consumer's problem graphically. For given initial m_t, the consumer's goal is to find the value of a such that Eq. (6)

Precautionary Saving and Precautionary Wealth, Fig. 1 Marginal utility of assets and of consumption



holds. The left-hand side of Eq. (6) is the upward-sloping locus. As for the two downward-sloping loci, the lower one reflects expected marginal value if the consumer is perfectly certain to receive the mean level of income $E_t\left[\tilde{y}_{t+1}\right]$, while the higher downward-sloping function corresponds to the case where income is uncertain.

When the risk is added, the optimal choice for end-of-period assets moves from a^* to a^{**}. Since $c_t = m_t - a_t$, the increase in a in response to risk corresponds to a reduction in consumption. This reduction in consumption is the precautionary saving induced by the risk.

For a given $v_{t+1}(m_{t+1})$, the exercise captured in the diagram can be conducted for every possible value of m_t, implicitly defining a consumption function $c_t(m_t)$.

Kimball (1990) shows that the index of absolute prudence $\frac{-v^m_{t+1}(m_{t+1})}{v^m_{t+1}(m_{t+1})}$ and the index of relative prudence $\frac{-v^m_{t+1}(m_{t+1})m_{t+1}}{v^m_{t+1}(m_{t+1})}$ are good measures of how much a risk of given size will shift the marginal value of assets curve $\omega'_t(a)$ to the right. For a constant relative risk aversion value function, relative prudence is equal to relative risk aversion plus 1. Kimball and Weil (2004) look at the strength of the precautionary saving motive when Kreps and Porteus (1978) preferences are used to break the usual equation $\varsigma = 1/\rho$ where ς is the elasticity of intertemporal substitution and ρ is relative risk aversion. In this more general case, the counterpart to relative prudence $\mathcal{P}$ is given by $\mathcal{P} = (1 + \varsigma\varepsilon)\rho$, where ε is the elasticity with which absolute risk aversion declines and absolute risk tolerance increases.

Note that, given the basic properties $\varsigma > 0$ and $\rho > 0$, a positive wealth elasticity of risk tolerance implies that $\mathcal{P} > \rho$. This is a special case of a much more general result first hinted at by Drèze and Modigliani (1972). Even for very exotic objective functions, the precautionary saving motive will always be stronger than risk aversion whenever ownership of more a_t due to a small forced reduction in consumption would lead an optimizing investor to bear more risk (a property that Drèze and Modigliani 1972 call ‘endogenously decreasing absolute risk aversion’). This general result holds because, if ownership of extra a_t due to a small forced reduction in consumption were to lead an optimizing investor to bear risks she was previously indifferent to, then reduced consumption must be complementary with bearing near-indifferent risks. The symmetry of complementarity then implies that, given a free choice of consumption levels, taking on an additional near-indifferent risk will lead an optimizing consumer to reduce consumption. For example, consider an agent with additive habit formation (as distinct from multiplicative habits, compare Carroll 2000), for whom reduced consumption not only increases assets but reduces the size of the consumption habit, and so unambiguously leads to more willingness to bear risks. Such an agent will want to reduce consumption if induced to take on an additional risk by a compensation that makes her indifferent to the risk. The size of the compensation is determined by risk aversion. Yet the compensation for the agent’s risk aversion is not enough to cancel out the precautionary saving effect of the risk.

Buffer Stock Wealth

The above discussion suggested that precautionary behaviour can be understood by considering a trade-off between the present (captured by $u(c_t)$) and the future (captured by $\omega_t(m_t - c_t)$).

That analysis was incomplete in a crucial respect: it took the initial level of resources, m_t, as given exogenously. But, arguably, the most important question about precautionary behaviour is how large an effect it has on the prevailing level of m. This cannot be answered using a framework that treats m as exogenous.

The framework can be extended to address this problem, by defining the problem in such a way that the functions v and ω reflect the discounted value of an infinite number of future periods. This is often accomplished by making assumptions under which optimal behaviour in every future period is identical to optimal behaviour in the current period; it is then possible to solve for a ‘consumption function’ that provides a complete characterization of the relationship between resources and spending.

The critical extra assumption is 'impatience', broadly construed as a condition on preferences that prevents wealth (or the wealth to income ratio) from growing to infinity. In the simplest version of the model where income does not grow, the required condition is $R\beta < 1$; for the appropriate condition in models with income growth, see Carroll (2004).

The exact nature of income risk turns out to be less important than the assumption of impatience. Here, we analyse a particularly simple case (which is an adaptation of a model by Toché 2005). There are two kinds of consumers: workers and retirees. Retirees have no labour income, and must live off their assets. Workers earn a fixed amount of labour income in each period, but face a constant danger of being exogenously forced into retirement. (Exogenous forced retirement is the sole source of risk in the model.)

Under these assumptions, if the utility function is of the standard constant relative risk aversion form $u(c) = c^{1-\rho}/(-\rho)$, optimal behaviour for retirees is very simple: they spend a constant fraction of m in each period, where the fraction depends on the degree of impatience and intertemporal substitution $(1/\rho)$.

The situation for workers is more interesting; it is depicted in Fig. 2. The simplest element of the figure is the line labelled 'Perm inc'. This shows, for any m, the level of spending that would leave expected m unchanged; it is equal to labour income plus the interest on capital income, and is upward sloping because a consumer with more m earns more capital income.

The assumption of impatience is reflected in the fact that the consumption function that would apply if uncertainty did not exist, $\overline{c}(m)$, is everywhere above the level of permanent income (income of the perfect-certainty consumer is adjusted downwards so that the reduction in unemployment risk does not cause an increase in mean income). In other words, an impatient consumer facing no uncertainty would choose to spend at a rate that cannot be sustained indefinitely.

The locus with arrows is the consumption function, which indicates the optimal level of spending (in the presence of uncertainty) for any given level of m. Since the difference between $c(m)$ and $\overline{c}(m)$ is purely the consequence of risk, that difference $\overline{c}(m) - c(m)$ constitutes the amount of precautionary saving associated with any specific m.

Standard assumptions about preferences and uncertainty imply that there will be an intersection between the permanent income locus and the consumption function. (For a proof that there will be only one intersection, see Carroll 2004.) The intersection defines a 'target' level for the buffer stock of wealth m: the level such that an employed consumer with this amount of resources today will end up with the same m next period. Dynamics are captured by the arrows, which indicate that, for initial values of m below the target, consumption is below permanent income, so m is increasing and consumption crawls upwards along the consumption function towards the target. For initial values of m above the target, consumption is above permanent income, so m is falling. The consumer holds a 'buffer stock' of wealth in an attempt to reach the 'target' level of wealth as defined above.

The existence of a target level of resources has many interesting implications. Perhaps the most surprising is that in long-run equilibrium the expected growth rate of consumption for employed consumers is unrelated to the interest rate or the degree of impatience.

To understand this point better, and to relate it to the literature, we restate it in a slightly more general form: The equilibrium expected growth rate of consumption for employed consumers is approximately equal to their predictable rate of income growth,

$$E_t\left[\Delta \log c^e_{t+1}\right] \approx g. \tag{7}$$

In many respects, the equilibrium equality of consumption growth and permanent income growth seems intuitive. However, it appears to conflict with a standard way of analysing consumption growth, which relies on the first-order condition from the optimization problem (the

'Euler equation'), which is often approximated by an equation of the form

$$E_t\left[\Delta \log c^e_{t+1}\right] \approx \rho^{-1}(r-\tau)+\varphi \tag{8}$$

where ρ is the coefficient of relative risk aversion and τ is the geometric rate at which future utility is discounted (related to the time preference factor β); φ is a term that reflects the contribution of precautionary motives to consumption growth.

The resolution of the apparent contradiction is that the precautionary component of consumption growth is endogenous; combining Eqs. (7) and (8) permits us to solve for the equilibrium value of the precautionary contribution to consumption growth:

$$\varphi \approx g-\rho^{-1}(r-\tau). \tag{9}$$

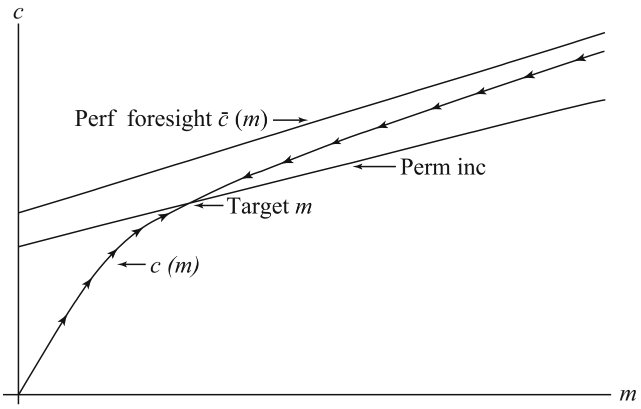
We return to this point below.

We can characterize the effect of uncertainty by noting three facts about Fig. 2: $c(m) < \bar{c}(m)$ (consumption is lower in the presence of uncertainty); $\lim_{m\to\infty}\bar{c}(m)c(m)=0$ (as wealth approaches infinity the effect of uncertainty in labour income vanishes); and $c(m)$ is strictly concave, so that the marginal propensity to consume out of a windfall increase in income, $c'(m)$, is greater for poor people than for rich people.

The concavity of the consumption function bears further comment. Intuitively, it can be understood in a similar light to the effect of liquidity constraints. A consumer who is subject to a currently binding liquidity constraint is someone for whom a marginal increase in cash will result in an immediate one-for-one increase in spending (a marginal propensity to consume, MPC, of 1). However, if the same consumer happened to have a large windfall transfer of cash (say, he wins the lottery), he would no longer be currently constrained, and his MPC would (presumably) be less than 1. In the case of precautionary saving, the ownership of an extra unit of wealth relaxes the suppression of consumption due to risk; this relaxation is more powerful for low-wealth consumers living on the edge of (precautionary) fear than for high-wealth consumers with plenty of resources. Thus, either liquidity constraints or precautionary motives or both will cause the consumption function to become concave (Carroll and Kimball 2005). Huggett (2004) shows that consumption concavity in turn implies greater equilibrium wealth.

Empirical evidence indicates that the wealth distribution is highly concentrated. This means that the owners of much of the aggregate capital stock probably inhabit the portion of the consumption function to the far right, where it approaches the linear consumption function that characterizes the perfect foresight solution. Note, however, that this does not necessarily imply that aggregate consumption behaviour will resemble that of a perfect foresight consumer, because a large proportion of aggregate consumption is accounted for by households with small amounts of market wealth. Spending of such households is probably determined much more by their permanent income than by their meagre wealth, and so it remains possible

Precautionary Saving and Precautionary Wealth, Fig. 2 The consumption function



that a high proportion of consumption is performed by households inhabiting the more nonlinear part of the consumption function.

Empirical Evidence

Euler Equation Methods

The early literature relevant to identifying the strength of precautionary motives tended to rely on Euler equation estimation (see Browning and Lusardi 1996 for a survey), often by estimating regression equations of the form

$$\Delta \log C_{t+1} = \alpha_0 + \alpha_1 E_t[r_{t+1}] \qquad (10)$$

and interpreting the coefficient on the interest rate term as an estimate of the inverse of the coefficient of relative risk aversion (CRAA) (which holds true under timeseparable CRRA utility, as in equation Eq. (8)). However, this analysis did not take into account the dependence of higher-order terms like φ on the independent variables (see Eq. (9)). Some papers like Dynan (1993) attempted to account for precautionary contributions to consumption growth; but see Carroll (2001) for a critique of the whole Euler equation literature (including the second-order approach).

Structural Estimation Using Micro Data

A new methodology for estimating the importance of precautionary motives was pioneered by Gourinchas and Parker (2002) and Cagetti (2003) (with a related earlier contribution by Palumbo 1999). Their idea was to calibrate an explicit life-cycle optimization problem using empirical data on the magnitude of household-level income shocks, and to search econometrically for the values of parameters such as the coefficient of relative risk aversion that maximized the model's ability to fit some measured feature of the empirical data. Gourinchas and Parker (2002) matched the profile of mean consumption over the lifetime; Cagetti (2003) matched the profile of median wealth. The intensity of the precautionary motive emerges, in each case, as an estimate of the coefficient of relative risk aversion, which Gourinchas and Parker (2002) put at about 1.4 and Cagetti (2003) finds to be somewhat larger (a value of 1 corresponds to logarithmic utility). One important caveat about these quantitative results is that the method's estimates of relative risk aversion depend on the model's assumption about the degree of risk households face. Recent work by Low, Meghir and Pistaferri (2005) that attempts to correct for measurement problems caused by job mobility suggest that the estimates of the magnitude of permanent shocks in Carroll and Samwick (1997) used for calibration by Gourinchas and Parker (2002) and Cagetti (2003) may be overstated by as much as 50 per cent. Re-estimation of the structural parameters using the Low et al. (2005) calibration would generate larger estimates of relative risk aversion.

Regression Evidence

A separate literature attempts direct empirical measurement of the relationship between uncertainty and wealth. To fix notation, index individual households by i and assume that uncertainty for household i in period t can be measured by some variable $\sigma_{t,i}$. Then in its simplest form the idea is to perform a regression of cash-on-hand on its determinants along the lines of

$$\log m_{t,i} = \sigma_{t,i}\gamma + Z_{t,i}\alpha + \varepsilon_{t,i} \qquad (11)$$

where Z is some set of variables that capture life cycle, time series, and other nonprecautionary effects. In principle, one can then calculate the predicted magnitude of m if everyone's uncertainty were set to zero (or some more sensible alternative like the minimum measured value of σ in the population).

This method permits the data to speak in a much less filtered way than the structural estimation approach. A drawback is that even if the magnitude of precautionary wealth could be estimated reliably and precisely, it would not be clear how to translate those estimates into a measure of relative risk aversion or some other set of behavioural parameters that could be used for analysing policy questions such as the optimal design of unemployment insurance or taxation.

A further disadvantage is that the method does not reliably yield the same answer in different data. Using a measure of subjective earnings uncertainty from a survey of Italian households, Guiso et al. (1992) estimate the precautionary component of wealth at only a few per cent, while Kazarosian (1997) and Carroll and Samwick (1998) estimate the precautionary component of wealth for typical US households to be in the range of 20–50 per cent. Hurst et al. (2005) argue that estimates of α are inordinately sensitive to whether business owners are included in the dataset; and work by Lusardi (1997, 1998) and Engen and Gruber (2001) implies much smaller precautionary wealth. Such large variation in empirical estimates is not plausibly attributable to actual behavioural differences across the various sample populations.

A problem that plagues all these efforts is identifying exogenous variations in uncertainty across households. The standard method has been to use patterns of variation across age, occupation, education, industry and other characteristics. This runs the danger that people who are more risk tolerant may both choose to work in a risky industry and choose not to save much, biasing downwards the estimate of the effect of an exogenous change in risk.

One recent paper attempts to get around this problem by using a natural experiment: Fuchs-Schündeln and Schündeln (2005) show that, before the collapse of the Berlin Wall, East German civil servants had similar income uncertainty to that faced by other East Germans. However, after the collapse of Communism, income uncertainty went up dramatically for most East Germans – but not for civil servants, who were given essentially the same risk-free jobs in the new merged government that they had had before the collapse. Fuchs-Schündeln and Schündeln (2005) show that, in accord with a model that includes substantial precautionary effects, saving rates of most East Germans increased sharply after unification, but saving rates of civil servants did not. By contrast, the West Germans – who would have been subject to more selection into jobs based on risk preferences – exhibited little difference in saving rates between civil servants and others with riskier jobs, either before or after reunification.

Survey Evidence

Given the difficulties of obtaining reliable quantitative measures of precautionary motives using the revealed preference econometric techniques sketched above, some researchers have turned to approaches that involve asking survey participants more direct questions.

Kennickell and Lusardi (2005) find that, when respondents for the 1995 and 1998 US *Survey of Consumer Finances* are asked their target level of precautionary wealth, most have little difficulty in answering the question: desired precautionary wealth represents about eight per cent of total net worth and 20 per cent of total financial wealth. They find that respondents cite a broad array of risks in making their precautionary targets: in addition to labour income risk, they face health risk, business risk, and the risk of unavoidable expenditures (such as home repairs). (Consumers are clearly aware of the theoretical point that a given dollar of wealth can provide self-insurance against multiple different kinds of risks, since the risks are not likely to be perfectly correlated with each other.)

Carefully designed survey questions can in principle also be used to elicit information on the strength of underlying preferences (like risk aversion) that determine precautionary behaviour. The principle that whenever risk-bearing increases with assets, the precautionary saving motive (prudence) must be stronger than risk aversion provides an important theoretical lower bound on the degree of prudence. Using survey responses to hypothetical gambles over lifetime income in the *Health and Retirement Study*, Kimballet al. (2005) estimate that relative risk aversion has a median of 6.3 and a mean of 8.2. (Note that because of Jensen's inequality, the mean of relative risk aversion $E\rho$ is larger than the reciprocal of the mean of relative risk tolerance $\frac{1}{E(1/\rho)}$.) These estimates of relative risk aversion imply precautionary saving motives much stronger than those that have been used empirically to match observed wealth holdings. This discrepancy remains unresolved.

Conclusion

The qualitative and quantitative aspects of the theory of precautionary behaviour are now well established. Less agreement exists about the strength of the precautionary saving motive and the magnitude of precautionary wealth. Structural models that match broad features of consumption and saving behaviour tend to produce estimates of the degree of prudence that are less than those obtained from theoretical models in combination with risk aversion estimates from survey evidence. Direct estimates of precautionary wealth seem to be sensitive to the exact empirical procedures used, and are subject to problems of unobserved heterogeneity. Thus, establishing the intensity of the precautionary saving motive and the magnitude of precautionary wealth remain lively areas of debate.

See Also

- ▶ Ambiguity and Ambiguity Aversion
- ▶ Elasticity of Intertemporal Substitution
- ▶ Euler Equations
- ▶ Intertemporal Choice
- ▶ Permanent-Income Hypothesis
- ▶ Time Preference

Bibliography

Barsky, R.B., N. Gregory Mankiw, and S.P. Zeldes. 1986. Ricardian consumers with Keynesian propensities. *American Economic Review* 76: 676–691.

Browning, M.J., and A. Lusardi. 1996. Household saving: Micro theories and micro facts. *Journal of Economic Literature* 34: 1797–1855.

Cagetti, M. 2003. Wealth accumulation over the life cycle and precautionary savings. *Journal of Business and Economic Statistics* 21: 339–353.

Carroll, C.D. 2000. Solving consumption models with multiplicative habits. *Economics Letters* 68: 67–77.

Carroll, C.D. 2001. Death to the log-linearized consumption Euler equation! (and very poor health to the second-order approximation). *Advances in Macroeconomics* 1, Article 6.

Carroll, C.D. 2004. Theoretical foundations of buffer stock saving. Working paper no. 10867. Cambridge, MA: NBER.

Carroll, C.D. and M. S. Kimball. 2005. Liquidity constraints and precautionary saving. Manuscript, Johns Hopkins University. Online. Availeable at: http://econ.jhu.edu/ccarroll/liquidRevised.pdf. Accessed 31 May 2007.

Carroll, C.D., and A.A. Samwick. 1997. The nature of precautionary wealth. *Journal of Monetary Economics* 40: 41–71.

Carroll, C.D., and A.A. Samwick. 1998. How important is precautionary saving? *The Review of Economics and Statistics* 80: 410–419.

Drèze, J.H., and F. Modigliani. 1972. Consumption decisions under uncertainty. *Journal of Economic Theory* 5: 308–335.

Dynan, K.E. 1993. How prudent are consumers? *Journal of Political Economy* 101: 1104–1113.

Engen, E., and J. Gruber. 2001. Unemployment insurance and precautionary saving. *Journal of Monetary Economics* 47: 545–579.

Fuchs-Schündeln, N., and M. Schündeln. 2005. Precautionary savings and selfselection: Evidence from the German reunification ‘experiment’. *Quarterly Journal of Economics* 120: 1085–1220.

Gourinchas, P.-O., and J. Parker. 2002. Consumption over the life cycle. *Econometrica* 70: 47–89.

Guiso, L., T. Jappelli, and D. Terlizzese. 1992. Earnings uncertainty and precautionary saving. *Journal of Monetary Economics* 302: 307–337.

Huggett, M. 2004. Precautionary wealth accumulation. *Review of Economic Studies* 71: 769–781.

Hurst, E., A. Lusardi, A. Kennickell and F. Torralba. 2005. Precautionary savings and the importance of business owners. NBER working paper no. 11731.

Kazarosian, M. 1997. Precautionary savings – a panel study. *The Review of Economics and Statistics* 79: 241–247.

Kennickell, A. and A. Lusardi 2005. Disentangling the importance of the precautionary saving motive. Working paper, Dartmouth College.

Kimball, M.S. 1990. Precautionary saving in the small and in the large. *Econometrica* 58: 53–73.

Kimball, M.S., C.R. Sahm, and M.D. Shapiro. 2005. *Imputing risk tolerance from survey responses*. Michigan: University of Michigan.

Kimball, M.S., and P. Weil. 2004. *Precautionary saving and consumption smoothing across time and possibilities*. Michigan: University of Michigan.

Kreps, D.M., and E.L. Porteus. 1978. Temporal resolution of uncertainty and dynamic choice theory. *Econometrica* 46: 185–200.

Laibson, D. 1997. Golden eggs and hyperbolic discounting. *Quarterly Journal of Economics* 112: 443–477.

Leland, H.E. 1968. Saving and uncertainty: The precautionary demand for saving. *Quarterly Journal of Economics* 82: 465–473.

Low, H., C. Meghir, and L. Pistaferri. 2005. *Wage risk and employment over the life cycle*. Stanford: Stanford University.

Lusardi, A. 1997. Precautionary saving and subjective earnings variance. *Economics Letters* 57: 319–326.
Lusardi, A. 1998. On the importance of the precautionary saving motive. *American Economic Review* 88: 449–453.
Miller, B.L. 1976. The effect on optimal consumption of increased uncertainty in labor income in the multiperiod case. *Journal of Economic Theory* 13: 154–167.
Palumbo, M.G. 1999. Uncertain medical expenses and precautionary saving near the end of the life cycle. *Review of Economic Studies* 66: 395–421.
Sibley, D.S. 1975. Permanent and transitory effects of optimal consumption with wage income uncertainty. *Journal of Economic Theory* 11: 68–82.
Toché, P. 2005. A tractable model of precautionary saving in continuous time. *Economics Letters* 87: 267–272.
Zeldes, S.P. 1989. Optimal consumption with stochastic income: Deviations from certainty equivalence. *Quarterly Journal of Economics* 104: 275–298.

Predator–Prey Models

R. M. Goodwin

Quite independently, two basic advances were made in the theory of animal populations by A.J. Lotka in 1925 and by Vito Volterra in 1926 and again in 1931. The resulting dynamical problem has implications for a wide range of cases, involving as it does, the possibilities of the extinction of one species, or the evolution to an equilibrated coexistence, or to a continuing oscillation. The problem was brought to Volterra by Ugo d'Ancona, who had found clear evidence for continued oscillation in the type of fish catches in the upper Adriatic. Volterra formulated and solved a system of two non-linear differential equations, in which the proportionate growth rate in each depended only on the level of the other. His solution is stable and positive, but is inappropriate for applied analysis, since it is structurally unstable. It is inappropriate in the sense that a mathematician may assume, as he does, a parameter to be exactly zero, but it is impermissible in an applied or empirical analysis like that of animal populations. In 1931 the Russian mathematician Kolmogoroff gave an elegant generalization which was both dynamically *and* structurally stable, yielding the three types of solution above, including a demonstration of a stable equilibrium motion, that is, a limit cycle. Various elaborations, along with qualitative analysis in phase space, are given in Hirsch and Smale (1974).

In spite of a potential relevance, these developments appear to have had no effect on economic theory until the appearance of a paper by Goodwin in 1965, and independently, two by Samuelson, in 1967 and again in 1971. In the latter the analysis is in terms of exploitation in conditions of diminishing returns, along with a limit cycle in connection with increasing returns. In the former, the symbiosis of workers and capitalists is presented in terms of conflict and yet mutual interdependence. Assuming steady growth of labour productivity and of labour force, a formulation in ratios results in a cycle of growth rates rather than levels. If wages are too high and profits too low, there is low growth and growing unemployment. Conversely, low wages and high profits bring high growth and falling unemployment. Too high and too low are defined in terms of producing that rate of accumulation necessary to keep the average growth rate of employment equal to that of the available labour force in such a way as to remain in the neighbourhood of full employment. Thus by dealing in distributive shares, the cycle includes growth and growth proceeds cyclically, a feature not present in the biological model. Consequently, though fluctuating, the economy produces a long-run, constant, average distribution of income, growth rate, and degree of unemployment. Even exogenous disturbances, leading to variations in initial conditions, do not alter the long-run averages, since they are independent of initial conditions and equal to equilibrium values. The model is susceptible to more realistic elaboration by the introduction of nominal and real wages, degree of capacity utilization, game theoretic formulation of wage bargaining, and more endogeneous treatment of productivity growth.

The model may be formulated in terms of the ratios, u, the share of wages in income, and *v*, the ratio of employment to available labour force. The result is the following matric equation:

$$\left\{\begin{matrix}\dot{u}/u\\ \dot{v}/v\end{matrix}\right\} = \begin{bmatrix}0 & +\alpha\\ -\rho & 0\end{bmatrix}\left\{\begin{matrix}u\\ v\end{matrix}\right\} + \left\{\begin{matrix}-\gamma\\ +\delta\end{matrix}\right\}$$
$$\times(\text{all parameters positive}).$$

Apart from the proportional growth rates, one sees the simplest, first order, linear differential equation, with the positive equilibria, $\bar{u} = \rho/\rho$ and $\bar{v} = \gamma/\alpha$. Because u depends positively on v and v negatively on u, it must oscillate, and because of the zeros on the diagonal it is stable dynamically and unstable structurally. Kolmogoroff wrote $\dot{u}/u = f(u,v)$ and $\dot{v}/v = g(u,v)$ giving an elegant, qualitative demonstration of all the possible types of motion, including an asymptotically stable limit cycle.

The system can be dazzlingly generalized from von Neumann's steady-state general equilibrium growth model. His turnpike, dynamic equilibrium path, being unstable, will alternately produce a high growth rate, rising real wage and decelerating growth, leading to low growth, low (relative to productivity) wage and rising growth rate. Hence the growth rate will vibrate about its equilibrium, never remain there, and will yield a constant, average, long-run growth rate.

Suitably modified this type of theory is of great importance for economists. It avoids the usual assumption that trend and cycle are independent, in the sense that each would exist in the absence of the other. It also is the natural way to treat technological progress, that is, an innovation in the form of either a new process of production or of a new good, is analogous to the introduction of a new species into a given ecological environment. The consequence is an explicable dynamical evolution to changing economic equilibria. This gives the essential view of capitalism as a permanent evolutionary process, or continuing morphogenesis.

See Also

- ▶ Bioeconomics
- ▶ Lotka, Alfred James (1880–1949)
- ▶ Volterra, Vito (1860–1940)

Bibliography

Goodwin, R.M. 1967. A growth cycle. In *Essays in economic dynamics*, ed. R.M. Goodwin. London: Macmillan, 1982.

Hirsch, M.W., and S. Smale. 1974. *Differential equations, dynamical systems, and linear algebra*. New York: Academic.

Kolmogoroff, A. 1931. On the theory of Volterra of the struggle for existence. *Journal of the Italian Actuaries*.

Lotka, A.J. 1925. *Elements of physical biology*. New York: Dover, 1956.

Samuelson, P.A. 1966. A universal cycle. In *The collected scientific papers of Paul A. Samuelson*, ed. J.E. Stiglitz. Cambridge, MA: MIT Press.

Samuelson, P.A. 1971. Generalized predator–prey oscillations in ecological and economic equilibrium. In *The collected scientific papers of Paul A. Samuelson*, vol. III, ed. R.C. Merton. Cambridge, MA: MIT Press, 1972.

Volterra, V. 1931. *Lectures on the mathematical theory of the struggle for existence*. Paris: Gauthier-Villars.

Predatory Pricing

Janusz A. Ordover

Abstract

Predatory pricing is a response to a rival that sacrifices part of the profit that could be earned under competitive circumstances were the rival to remain viable, in order to lessen competition and gain consequent monopoly profit. The presence of intertemporal cost and/or demand linkages as well as network effects complicates the formulation of pricing rules that would distinguish legitimate from exclusionary pricing behaviour, and suggests that standard (non-strategic) ▶ models of markets do not necessarily offer much help in gauging the rationality of predation.

Keywords

Above-cost pricing; Antitrust policies; Barriers to entry; Chain-store paradox; Entry; Exit; Incomplete information; Increasing returns; Intertemporal scope economies; Marginal and average cost pricing; Natural monopoly; Network goods; Predatory pricing; Returns to scale; Standardization; Two-sided platforms

JEL-Classifications
D4

Although neither courts nor legal and economic scholars agree on a broad definition of predatory behaviour, the minimal consensus (if such exists) is that predatory pricing entails selling a product ‘below cost’ in order to induce a rival’s exit, or deter future entry or competition. More broadly, ‘predatory behaviour is a response to a rival that sacrifices part of the profit that could be earned under competitive circumstances, were the rival to remain viable, in order to [lessen competition] and gain consequent monopoly profit’ (Ordover and Willig 1981, pp. 9–10).

The broader definition is necessary, at least in part because in many market scenarios, comparisons of prices to marginal cost offer little guidance as to what constitutes competitive, as opposed to predatory, pricing. A multi-product firm might offer one of a pair of complementary products at a price above incremental cost, and yet still be engaged in predation if the price at which it offers the pair as a bundle is sufficiently higher than the incremental cost of the second component (see, for example, Baumol and Sidak 1994, ch. 7). Another scenario is markets with intertemporal scope economies, as in Cabral and Riordan (1994), in which two firms race to exploit learning economies or establish their respective products as industry standard. In such markets, it may be profitable for the ‘leading’ firm to price below cost in order to induce the rival’s exit. Discouraging such pricing would damage competition for the market for the sake of protecting competition in the market.

Similar issues arise in markets of network goods, where the product is more valuable to a user the more other people use it (for example, fax machines). These markets are characterized by increasing returns, and may be subject to a ‘tipping’ point at which one firm achieves natural monopoly, which is an efficient outcome since it increases consumer welfare by increasing network benefits through standardization. Farrell and Katz (2005) find that although rules to prevent predation, such as the Ordover–Willig rule, can improve welfare, they can also harm it in network markets ‘by preventing firms from internalizing the benefits of increasing returns to scale’.

In these cases, aggressive pricing is not designed to drive the rival into bankruptcy, but to make it realize that the ‘game’ is over from a strategic standpoint. When the market participants jockey for market leadership, pricing below short-run marginal cost (SRMC) could be a rational, non-predatory strategy. Evans and Schmalensee (2002) go so far as to advocate an approach under which, ‘if a defendant can establish that the relevant market is characterized by winner-take-all competition, then they have provided a complete defense against a charge of predatory behaviour’.

Another setting in which simple pricing rules can lead to wrong inferences involves pricing by so-called two-sided platforms, intermediaries that link two distinct groups of customers (for example, Rochet and Tirole 2003, 2006; Armstrong 2007). Such intermediaries frequently subsidize customers on one side in order to induce the other side to join the platform and enhance the value to all participants. Thus, below-cost pricing is compensated by above-cost pricing on the other side and by its impact on the overall level of activity on the platform. Such pricing may, of course, harm rivals who only operate on one side of the platform.

The presence of intertemporal cost and/or demand linkages as well as network effects of various kinds complicates the formulation of pricing rules that would sort out legitimate from exclusionary pricing behaviour. It also suggests that standard (non-strategic) models of markets do not necessarily offer much help in gauging the rationality of predation.

The Apparent Irrationality of Predatory Pricing

The Chicago School critique of traditional views of predatory pricing rested on the hypothesis that losses sustained during the predatory campaign will ordinarily exceed the more speculative gains from attempted supercompetitive pricing following the elimination of the prey. In his examination of *Standard Oil of New Jersey v. US* case, McGee (1958) pointed out that, in order for the predator to succeed in driving out an equally efficient rival, it must be prepared to serve the whole market by itself at an unremunerative price, while the prey can temporarily shut down its operations and restart them during the recoupment phase. Moreover, even if the prey exists permanently, productive assets may remain and could be purchased at scrap value by an opportunistic buyer. Easterbrook (1981) further observed that customers might protect themselves against post-predation exploitation by keeping the prey in business, even if the product is available from the predator at a lower price, thereby denying the predator an opportunity to drive the rival out. And the prey may also be financed by lenders who (correctly) anticipate that, once the predator gives up, additional profits will be generated with which to repay the loan. McGee also noted that it is generally cheaper to purchase the rival rather than to prey on it. Hence, according to McGee, even if feasible, predation is irrational.

There are several problems with McGee's merger argument: buying a single rival may induce others to enter solely to be bought out at a premium (Rasmusen 1985); the acquisition price itself may depend on the predator's established reputation for aggressive pricing (Burns 1986; Saloner 1987); there may be legal constraints on mergers so that when it is most advantageous, it is also likely to violate anti-merger legislation (Posner 1976).

McGee's critique significantly influenced antitrust policies regulating pricing conduct, but stopped well short of offering a rigorous model in which predation was irrational. Selten's (1978) *chain store paradox* does so. Intuition suggests that an incumbent operating in a sequence of markets, each with an entrant, may predate in the first few markets to establish a 'reputation' for toughness and thereby deter the remaining entrants. The intuition fails, however, by 'backward induction': the entrant in the last market will correctly disregard the incumbent's behaviour in the preceding markets and conclude that its entry will be accommodated because, with no reputation to be concerned about, the incumbent has no reason to predate. The penultimate entrant reasons additionally that its predation will not deter entry into the next, and so it too enters, expecting to be accommodated. Inexorable logic leads to the conclusion that the incumbent will not predate and entry will occur in all the markets (see Ordover and Saloner 1989; Phlips 1995).

Economic Models of Rational Predation

In settings that dispense with some of Selten's assumptions, predation can emerge as a rational strategy.

The Long Purse

One typical predatory pricing story involves an incumbent with a 'deep pocket', who by pricing aggressively can drive out a financially constrained rival (see Telser 1966). In order to induce exit, the incumbent drops the price to the *rival's (not necessarily its own) variable cost*. The rival, who also incurs fixed costs, soon exhausts its financial resources and leaves the market, enabling the incumbent to raise its price to monopoly level to recoup the costs of the predatory campaign.

Here, the mere threat of rational predation drives the opponent out. Clearly, a rational rival should leave at the first indication of predation, and not squander resources when exit is inevitable. In fact, rational firms with limited resources ought to stay out of a market occupied by an incumbent with a long purse. The 'long purse' story is not an entirely plausible basis for rational predation, but rather of entry deterrence, via a *credible threat* of post-entry predation – indeed, a costless one, as Benoit (1984) shows.

The long-purse model also ignores the possibility of profit-seeking investors financing the preyed-upon firm, in order to extend its purse. In Bolton and Scharfstein (1990) and Fudenberg and Tirole (1986), the predator imposes losses on its prey in order to signal to investors that the prey is financially troubled. Even when everyone knows it is profitable for the rival to remain in the industry, Bolton and Scharfstein argue that financial market predation induces exit because agency problems in financial contracting mean that reducing the sensitivity of the refinancing decision to the firm's performance exacerbates managerial incentive problems.

Predation for Reputation

Other models operate by making 'predation for reputation' a *rational strategy* (see Ordover and Saloner 1989; Milgrom and Roberts 1990; Phlips 1995, for more detailed analyses). For example, the game may have no 'end' from which to reason backwards (see Milgrom and Roberts 1982b). Or there may be incomplete information, with different incumbent 'types', as in the seminal papers by Kreps and Wilson (1982), Milgrom and Roberts (1982b), and Kreps et al. (1982). A 'weak' incumbent, who would otherwise prefer to share a market, can falsely establish a 'tough' reputation by fighting at the first opportunity, and so convince all possible future entrants of its toughness and deters future entry, since if every incumbent were to predate, the reputational value of fighting would be dissipated and entry would occur, the probability of equilibrium predation must be positive, but less than one. This predatory story can be enriched in several ways: see, for example, Milgrom and Roberts (1982b) and Easley et al. (1985), whose work is reviewed in Phlips (1995).

Signalling Predation

Under imperfect information, predation can also be used to induce the rival's exit. For example, the rival may not have perfect knowledge of the incumbent's costs or its new product's demand. In these plausible market settings, the better-informed incumbent may price low in order to *signal* to the rival that exiting the market is preferable to staying (see for example, Milgrom and Roberts 1982a). Even if low pricing does not deter entry, it may convince the rival to curtail its competitive ardour. Or, as Saloner (1987) shows, turning McGee's merger argument on its head, it may improve the terms of a buyout offer, by convincing the quarry that accepting a cheap offer is preferable to sharing a market with a low-cost competitor.

Signalling predation will be especially effective when the rival firm tries to gauge a new product's profitability from its reception in the 'test market'. Firms with competing products will wish to 'jam' the signal (see Salop and Shapiro 1980; Scharfstein 1984; Roberts 1986).

Empirical Evidence

Recent empirical work has supported the rational predation models. A broad survey found that predatory pricing was present in 27 of 40 litigated cases in which the legal record was sufficiently informative (Zerbe and Mumford 1996). In addition, several case studies, taken collectively, provide evidence that dominant firms have engaged in predatory behaviour, thereby undermining the Chicago School's claims about its irrationality.

Weiman and Levin (1994) provide evidence of predatory behaviour by Southern Bell Telephone Company from 1894 to 1912 when independent phone companies threatened entry. Genesove and Mullin (2006) provide evidence of predatory behaviour in the American sugar refining industry before the First World War. They show that the price wars following two episodes of entry were predatory by comparing price to marginal costs and by constructing predicted competitive cost margins that they show to exceed observed margins. Granitz and Klein (1996) re-examine McGee's Standard Oil case and find evidence that Standard had in fact acted as a predator, by threatening to withhold crude shipments from any railroad that did not participate in a railroad cartel, in return receiving discounted shipping rates that left Standard's competitors to sell out at depressed prices.

Von Hohenbalken and West (1984) and West and Von Hohenbalken (1984) provide evidence that a leading Canadian supermarket chain engaged in a predatory location strategy. In a subsequent study (1986), they show that the strategy also gave the chain a reputation for aggressive pricing that deterred future entrants, thereby supporting the reputation model of Kreps and Wilson (1982) and Milgrom and Roberts (1982b), among others.

Burns (1986) similarly finds systematic empirical evidence supporting the theory that firms can acquire a reputation for following through on predatory commitments from past predatory behaviour. He finds that the American Tobacco Company from 1891 to 1906 set up bogus independents that it secretly controlled to sell at low prices in the prey's territories, thereby allowing the predator to maintain its monopoly by acquiring the assets of the prey, as well as other competitors not yet preyed upon, at artificially low prices. This study lends considerable credence to the view that predation can improve the terms of a takeover.

In contrast, Lott (1999) argues on the basis of an empirical survey of firms accused of predation between 1963 and 1982, that such firms did not have the necessary contractual and non-contractual arrangements to provide managers with incentives to engage in costly predatory behaviour, which should be necessary to lend credibility to the strategies. This critique is quite powerful but it goes deeper than just an attack on the credibility of predation. It suggests that, unless the principals (owners) can induce agents (managers) to forgo current profits for the sake of any future monopoly profits, managers may simply decide not to implement such strategies. On the other hand, even the most casual empiricism also suggests that such obstacles need not be insurmountable.

Legal Tests for Price Predation

Price predation is easily confused with intense competition. Sharp demarcation lines are particularly difficult in strategic environments in which price predation could prove profitable. What should the public policy response be to the inherent difficulties in formulating antitrust rules governing dominant firm pricing? The legal–economic literature offers three distinct responses.

At one extreme, the Chicago School has urged removal of virtually all constraints on single firm pricing behaviour (as well as other forms of unilateral conduct). The rationale is simple: firms should not be discouraged from aggressively competing for and protecting their market positions. Further, since markets are quickly self-correcting (unless protected by governmental grants of monopoly), marketplace advantages unrelated to superior skill and efficiency are quickly driven away by competition. Consequently, anti-competitive conduct – including price predation – is, in general, irrational, and any attempts to forbid such conduct are likely to do more harm than good.

At the other extreme lies an open-ended, rule-of-reason analysis without any specific rules (see, for example, Scherer 1976; Comanor and Frech 1984). There are serious problems with this approach, however. First, it is not clear whether it can be implemented effectively in the adversarial setting of antitrust litigation. Second, because it offers no standards for what constitutes lawful conduct, this approach complicates business planning and may increase incentives for the abuse of antitrust laws.

The third public policy response is consistent with most legal–economic commentary and judicial practice. Although the US courts have rarely found price predation, they have been unwilling to rule it out completely. Instead, the courts have adopted a set of 'filters' designed to screen potentially meritorious claims of anticompetitive pricing conduct from those that are probably without merit (see, for example, Joskow and Klevorick 1979; Easterbrook 1984; Baker 1994; Elzinga and Mills 1994). The rest of this section discusses these filters, first reviewing proposed direct tests for predatory pricing and then addressing the question of 'recoupment' as a precondition for a finding of price predation.

Pricing Tests

The Areeda–Turner Test (Areeda and Turner 1975, 1978)

Areeda and Turner proposed that any price above 'reasonably anticipated' SRMC should be lawful, and any below, deemed predatory. US courts rapidly embraced this test, which is now a dominant test (see, for example, Areeda and Hovenkamp 1993; Denger and Herfort 1994; Green et al. 1996).

Because SRMC is difficult to estimate, Areeda and Turner recommend the average variable cost (AVC) as a workable surrogate. However, AVC is a good surrogate only when it does not diverge significantly from SRMC. Indeed, Areeda and Turner's analysis of the appropriate measures of AVC is inadequate because it does not derive correct cost concepts from the analysis of the predatory conduct itself and, consequently, fails to provide adequate guidance on how to treat such important components of costs as capital and advertising expenses (see Ordover and Willig 1981; Baumol 1996). The main problem with the Areeda–Turner test, however, is that it is based on an analysis of a firm's behaviour in a market situation – a temporary price cut by a single-product single-market firm – in which profitable predation is unlikely.

The Areeda-Turner Paper

The Areeda and Turner paper generated a flow of alternative tests. For example, the *Joskow–Klevorick test* (1979) offers a two-tier test for price predation. The first step examines whether the structural preconditions for successful and rational predation exist. Because the first step eliminates many baseless claims, Joskow and Klevorick tighten the price comparison in the second step, and propose that any price below average *total* cost be presumptively illegal. The rationale is that in a competitive market, the equilibrium, long-run price will equal AVC and that, furthermore, it is unlikely that a post-entry price in a market predisposed to predation would be so low as to impose losses on the incumbent dominant firm. Some courts have used the Joskow–Klevorick test as an alternative to the Areeda–Turner test, especially when entry barriers are high. Moreover, an analysis of structural and other requirements for rational predation is now central to the analysis of a predatory pricing case.

Williamson (1977) and Baumol (1979) propose tests that isolate the *strategic* aspects of the incumbent's responses to entry. Both would condemn 'window shade'- type behaviour by the incumbent, that is, low price (high output) when the rival is in, followed by high price (low output) when the rival is out, and require that the dominant incumbent stick with its aggressive strategy for a prescribed period of time. These tests have not, however, been adopted by the courts.

Both these proposals can be criticized on various grounds (see Ordover and Saloner 1989; Phlips 1995). Areeda and Hovenkamp (1993) offer a spirited defence of the original Areeda–Turner rule against its critics and review the alternatives, which they find less desirable than the Areeda–Turner rule.

Above-Cost Predation and the Edlin Test

In recent years a debate has ensued whether 'above-cost' pricing can also be predatory. In 1999, for example, the United States sued American Airlines on the theory that it was predatory to respond to entry with business practices that, even if above cost, 'clearly' sacrificed profits because it allegedly shifted airplanes from profitable routes to routes on which it was fighting the low-cost carriers. Edlin (2002) supports the move to prohibit above-cost predation because '[a]n incumbent monopoly with a significant cost or noncost advantage over entrants . . . can use these advantages to drive entrants from the market by pricing below their cost, but above its own' and proposes a rule that would prohibit an incumbent monopoly, when faced with an entrant charging at least 20 per cent below the prevailing price, from cutting its own prices for 12–18 months or until it loses its monopoly position.

According to Edlin, this rule means that matching competitors' prices after entry is no longer a cheap substitute for actually charging low prices in the first place, so consumers benefit. He explains that existing predation law means that the incumbent will not lower prices until there is an entrant and, since the potential entrant will not, in fact, enter, consumers always pay high prices to the incumbent. The predatory problem, he explains, occurs not after exit, but before entry. His rule, he argues, would address the problem by allowing firms that would otherwise fear being driven out of the market with above-cost predation to enter profitably, and it would benefit consumers because incumbents would charge lower prices to limit entry and because there would be more entry of competitors.

P

Elhauge (2003) responds that an above-cost pricing rule is ill-advised for three reasons: (*a*) it can often penalize efficient pricing behaviour when incumbents do not even have the market power to restrict output, for example when above-cost price cuts are an efficient response to deviations from the output-maximizing price-discrimination schedule in competitive markets; (*b*) it has mainly undesirable effects, such as raising post-entry prices and harming consumer welfare when the entrant is less efficient than the incumbent; and (*c*) it suffers from unavoidable implementation difficulties, such as ascertaining the moment of entry, dealing with quality changes designed to evade the restriction, and defining a post-entry price floor that will cause inefficiencies. He argues that part of the reason for the debate about whether to expand predation to above-cost pricing is ambiguity over the definition of 'costs' in the legal tests. He concludes that costs should be defined functionally as whichever cost measure assures that prices above costs cannot deter or drive out equally efficient rivals, a definition which he argues would resolve apparent anomalies in current predatory pricing law. Of course, from the standpoint of basic economics, it is always the 'opportunity cost' that provides the right measure of cost to be used. But this may be too much for the courts as calculations of opportunity costs are far from simple.

The Recoupment Test

The recoupment test is a potentially useful step in a summary judgement proceeding because it enables the fact-finder to dismiss allegations of predation without engaging in an extensive (and time-consuming) investigation of price-cost margins and other indicia of predatory *conduct*. On the other hand, the evidence that price is below the pertinent cost floor should perhaps obviate the need to enquire whether recoupment is feasible or not: the firm's conduct *reveals* its belief that recoupment is possible. In essence, the recoupment test substitutes the court's assessment of the likelihood of success for the independent business judgement of the alleged predator. Likewise, Hemphil (2001) argues that the recoupment analysis should not consider the firm's conduct at all, but rather should limit itself to an analysis of the structural features of the market, such as asymmetric information and linkages across markets, that might deter entry and allow the predator to profit from its ill-gotten monopoly once the competitor has been eliminated (see also Ordover and Willig 1981). The recoupment test has also been criticized by Edlin and Farrell (2004) on the grounds that quantifying how the predator might benefit from its behaviour is difficult; courts should thus pay more attention, they argue, to serious consumer harm or harm to economic efficiency – 'recoupment as harm' rather than 'recoupment as reality check'.

Critiques of the Current Legal Test of Predation

The Supreme Court's two-prong test in *Brooke Group* (1993) (price-cost and recoupment) created a high burden of proof for plaintiffs that solidified the Court's embrace of the Chicago School view that predatory pricing is 'rarely tried, and even more rarely successful'. In the ensuing six years after *Brooke Group*, plaintiffs had not won a single predatory pricing case in federal court and, more striking, all but three of 39 reported decisions were dismissed or failed to survive summary judgment (see Brodley et al. 2000).

Brodley et al. (2000) criticize the courts for adhering to this 'static, non-strategic view of predatory pricing' at the same time that modern economic theory and empirical evidence have demonstrated the prevalence of predatory pricing. Based on this modern strategic theory, they propose a legal rule that, they argue, would augment existing practice in two critical respects: (1) it would explicitly permit proof of predation based on modern economics, and (2) it would expand the standard efficiencies and business justification defences to encompass pro-competitive dynamic gains, such as the learning-by-doing and network markets discussed earlier.

In reply, Elzinga and Mills (2001) fault the proposal for ignoring that predatory pricing is in practice very uncommon. They also argue that such theory lacks factual support and is not yet well developed enough to incorporate into antitrust rules. As a result, to permit predation to be proven by reference to modern strategic theory risks over-enforcement. (Bolton et al. 2000, respond by noting that the heavy factual burdens on the defendant and fully developed efficiencies defence available to the defendant mitigate the risk of over-enforcement.)

Marx and Shaffer (1999) argue that, for intermediate goods markets, the Supreme Court's two-prong test in *Brooke Group* may be over-inclusive, as low-cost pricing and recoupment can both occur with the rival supplier, although harmed, remaining in the market, and welfare actually increasing (because the increase in consumer surplus outweighs the reduction in overall joint profit associated with the pricing distortion).

Conclusion

The three decades of research on predatory pricing since Areeda and Turner (1975) lead to the following policy lessons.

First, the strategic approach to modelling pricing debunked the comfortable position that predation is more costly to predator than prey, and hence irrational and unlikely to occur.

Second, given the non-competitive structure and asymmetries of information in the relevant markets, there is no bright line standard for predatory pricing that both proscribes pricing behaviour that reduces economic welfare and does not discourage pro-competitive behaviour.

Third, the focus on price predation to the exclusion of other types of business conduct seems misplaced, given the richness of strategies used by firms in their battles for market share (Ordover and Willig 1981). Many of these strategies are likely to be more successful than price predation in inducing the exit of efficient rivals, and do not require sustained periods of losses.

Fourth, the courts' shift from vague inquiries into the 'intent' of the alleged predator's actions to more rigorous price and cost comparisons and assessments of the likelihood of recoupment has not benefited plaintiffs in predatory pricing litigation.

See Also

- ▶ Game Theory
- ▶ Monopoly

Bibliography

Areeda, P.E., and H. Hovenkamp. 1986. *Antitrust law: 1986 supplement*. Boston: Little, Brown.

Areeda, P.E., and H. Hovenkamp. 1993. *Antitrust law (supplement)*. Boston: Little, Brown.

Areeda, P.E., and D.F. Turner. 1975. Predatory pricing and related practices under section 2 of the Sherman Act. *Harvard Law Review* 88: 697–733.

Areeda, P.E., and D.F. Turner. 1978. *Antitrust law*. Boston: Little Brown.

Armstrong, M. 2007. Two-sided markets: Economic theory and policy implications. In *Recent developments in antitrust*, ed. J.P. Choi. Cambridge, MA: MIT Press.

Baker, J.B. 1994. Predatory pricing after *Brooke Group:* An economic perspective. *Antitrust Law Journal* 64: 585–604.

Baumol, W.J. 1979. Quasi-permanence of price reductions: A policy for prevention of predatory pricing. *The Yale Law Journal* 89: 1–26.

Baumol, W.J. 1996. Predation and the logic of the average variable cost test. *Journal of Law and Economics* 39: 49–72.

Baumol, W.J., J.C. Panzar, and R.D. Willig. 1982. *Contestable markets and the theory of industry structure*. New York: Harcourt, Brace Jovanovich.

Baumol, W.J., and J.G. Sidak. 1994. *Toward competition in local telephony*. Cambridge, MA: MIT Press.

Benoit, J.P. 1984. Financially constrained entry in a game with incomplete information. *RAND Journal of Economics* 15: 490–499.

Bolton, P., J.F. Brodley, and M.H. Riordan. 2000. Predatory pricing: Strategic theory and legal policy. *Georgetown Law Journal* 88: 2239–2330.

Bolton, P., and D. Scharfstein. 1990. A theory of predation based on agency problems in financial contracting. *American Economic Review* 80: 93–106.

Bork, R. 1978. *Antitrust paradox*. New York: Basic Books.

Brodley, J.F., P. Bolton, and M.H. Riordan. 2000. Predatory pricing: Strategic theory and legal policy. *Georgetown Law Journal* 88: 2239.

Brooke Group Led v. Brown & Williamson Tobacco Corp*, 113 S Ct. 2578 (*1993*).*

Burns, M.R. 1986. Predatory pricing and the acquisition cost of competitors. *Journal of Political Economy* 94: 266–296.

Cabral, L.M.B., and M.H. Riordan. 1994. The learning curve, market dominance, and predatory pricing. *Econometrica* 62: 115–140.

Comanor, W.S., and H.E. Frech III. 1984. Strategic behavior and antitrust analysis. *American Economic Review* 74: 372–376.

Denger, M.L., and J.A. Herfort. 1994. Predatory pricing claims after *Brooke Group*. *Antitrust Law Journal* 62: 541–558.

Easley, D., R.T. Masson, and R.J. Reynolds. 1985. Preying for time. *Journal of Industrial Organization* 33: 445–460.

Easterbrook, F.H. 1981. Predatory strategies and counterstrategies. *University of Chicago Law Review* 48: 263–337.

Easterbrook, F.H. 1984. The limits of antitrust. *Texas Law Review* 63: 1–40.

Edlin, A. 2002. Stopping above-cost predatory pricing. *The Yale Law Journal* 111: 941–991.

Edlin, A., and J. Farrell. 2004. The American airlines case: A chance to clarify predation policy. In *The antitrust revolution*, ed. J.E. Kwoka Jr. and L.J. White, 4th ed. Oxford: Oxford University Press.

Elhauge, E. 2003. Why above-cost price cuts to drive out entrants are not predatory: And the implications for defining costs and market power. *The Yale Law Journal* 112: 681–827.

Elzinga, K.G., and D.E. Mills. 1994. Trumping the Areeda–Turner test the recoupment standard in *Brooke Group*. *Antitrust Law Journal* 62: 559–584.

Evans, D.S., and R. Schmalensee. 2002. Some economic aspects of antitrust analysis in dynamically competitive industries. In *Innovation policy and the economy*, ed. A.B.. Jae, J. Lerner, and S. Stern, vol. 2. Cambridge, MA: NBER and MIT Press.

Farrell, J., and M. Katz. 2005. Competition or predation? Consumer coordination, strategic pricing and price floors in network markets. *Journal of Industrial Economics* 53: 203–231.

Fudenberg, D., and J. Tirole. 1986. A 'signal-jamming' theory of predation. *RAND Journal of Economics* 17: 366–376.

Genesove, D., and W. Mullin. 2006. Predation and its rate of return: The sugar industry, 1887–1914. *RAND Journal of Economics* 17: 366–376.

Granitz, E., and B. Klein. 1996. Monopolization by 'raising rivals' costs': The standard oil case. *Journal of Law and Economics* 39: 1–47.

Green, W., and J.A. Ordover. 1996. *Predatory pricing* . Chicago: American Bar Association.Monograph No. 22

Green, W., et al. 1996. *Predatory pricing*. Chicago: American Bar Association.

Hemphil, S. 2001. Note, the role of recoupment in predatory pricing analysis. *Stanford Law Review* 53: 1581–1612.

Joskow, P.L., and A.K. Klevorick. 1979. A framework for analyzing predatory pricing policy. *The Yale Law Journal* 89: 213–270.

Kreps, D., P. Milgrom, J. Roberts, and R. Wilson. 1982. Rational cooperation in the finitely-repeated Prisoner's Dilemma. *Journal of Economic Theory* 27: 245–252.

Kreps, D., and R. Wilson. 1982. Reputation and imperfect information. *Journal of Economic Theory* 27: 253–279.

Lott, J. Jr. 1999. *Are predatory commitments credible?* Chicago: University of Chicago Press.

Marx, L.M., and G. Shaffer. 1999. Predatory accommodation: Below-cost pricing without exclusion in intermediate goods markets. *RAND Journal of Economics* 30: 22–43.

McGee, J. 1958. Predatory price cutting: The *Standard Oil* (NJ) case. *Journal of Law and Economics* 1: 137–169.

McGee, J. 1980. Predatory pricing revisited. *Journal of Law and Economics* 23: 289–330.

Milgrom, P., and J. Roberts. 1982a. Limit pricing and entry under incomplete information: An equilibrium analysis. *Econometrica* 50: 443–459.

Milgrom, P., and J. Roberts. 1982b. Predation, reputation and entry deterrence. *Journal of Economic Theory* 27: 280–312.

Milgrom, P., and J. Roberts. 1990. The new theories of predatory pricing. In *Industrial structure in the new industrial economics*, ed. G. Bonnano and D. Brandolini. Oxford: Clarendon Press.

Elzinga, K.G., and D.E. Mills. 2001. Predatory pricing and strategic theory. *Georgetown Law Journal* 89: 2475–2493.

Ordover, J.A., and G. Saloner. 1989. Predation, monopolization and antitrust. In *Handbook of industrial organization*, ed. R. Schmalensee and R.D. Willig. New York: North-Holland.

Ordover, J.A., and R.D. Willig. 1981. An economic definition of predation: Pricing and product innovation. *The Yale Law Journal* 91: 8–53.

Ordover, J.A., and R.D. Willig. 1995. Economists' view: The Department of Justice draft guidelines for the licensing and acquisition of intellectual property. *Antitrust Magazine* 9: 29–36.

Phlips, L. 1995. *Competition policy: A game-theoretic perspective*. Cambridge: Cambridge University Press.

Posner, R. 1976. Predatory pricing. In *Antitrust law: An economic perspective*, ed. R. Posner. Chicago: University of Chicago Press.

Rasmusen, E. 1985. *Entry for buyout*. Working paper, Graduate School of Management, UCLA.

Roberts, J. 1986. A signalling model of predatory pricing. *Oxford Economic Papers (Suppl.) NS* 38: 75–93.

Rochet, J-Ch., and J. Tirole. 2003. Platform competition in two-sided markets. *Journal of the European Economic Association* 1: 990–1029.

Rochet, J-Ch., and J. Tirole. 2006. Two-sided markets: A progress report. *RAND Journal of Economics* 21: 172–187.

Saloner, G. 1987. Predation, mergers and incomplete information. *RAND Journal of Economics* 18: 165–186.

Salop, S.C. and Shapiro, C. 1980. *A guide to test market predation*. Working paper, Federal Trade Commission, Bureau of Economics.

Scharfstein, O. 1984. A policy to prevent rational test-market predation. *RAND Journal of Economics* 15: 229–243.

Scherer, F.M. 1970. *Industrial markets structure and economic performance.* Chicago: Rand McNally; 2nd edn, 1980.

Scherer, F.M. 1976. Predatory pricing and the Sherman Act: A comment. *Harvard Law Review* 89: 869–890.

Selten, R. 1978. The chain-store paradox. *Theory and Decision* 9: 127–159.

Standard Oil Co. of New Jersey v. US*, 221 US 1* (1911).

Telser, L.G. 1966. Cutthroat competition and the long purse. *Journal of Law and Economics* 9: 259–277.

Von Hohenbalken, B., and D. West. 1984. Predation among supermarkets: An algorithmic locational analysis. *Journal of Urban Economics* 15: 244–257.

Von Hohenbalken, B., and D. West. 1986. Empirical tests for predatory reputation. *Canadian Journal of Economics* 19: 160–178.

Weiman, D., and R. Levin. 1994. Preying for monopoly? The case of Southern Bell Telephone Company, 1894–1912. *Journal of Political Economy* 102: 103–126.

West, D., and B. Van Hohenbalken. 1984. Spatial predation in a Canadian retail oligopoly. *Journal of Regional Science* 24: 415–429.

Williamson, O.E. 1977. Predatory pricing: A strategic and welfare analysis. *The Yale Law Journal* 87: 284–340.

Zerbe, R. Jr., and M. Mumford. 1996. Does predatory pricing exist? Economic theory and the courts after *Brooke Group*. *Antitrust Bulletin* 41: 949–985.

Prediction

P. Whittle

Any rational theory of prediction must be based upon a model. Enoch Powell expresses this view when he says, ‘The prophets were not soothsayers; they were expounders.’

We shall formulate models in discrete time, so that the time variable *t* can be assumed to take integral values. The value of a variable *x* at time *t* will be denoted x_t. We shall frequently denote the observation taken at time *t* by y_t (usually vector-valued) and shall then denote the *observation history*

$$(y_t, y_{t-1}, \ldots)$$

available at time *t* by Y_t. The estimate of a quantity *u* based upon Y_t will be denoted $u^{(t)}$. Thus $x_{t+m}^{(t)}$ is, for positive *m*, the predictor of x_{t+m} formed at time *t*. The linear *linear least square* (LLS) criterion chooses $u^{(t)}$ as the linear function of Y_t that minimizes the mean square deviation $E[u\text{–}u^{(t)}]^2$ (or a matrix analogue if *u* is vector-valued). If all variables are jointly normally distributed (*Gaussian*, henceforth), then this $u^{(t)}$ can also be characterized as the conditional expectation $E[u|Y_t]$ or as the maximum likelihood (ML) estimate of *u* for given Y_t.

There are two techniques useful in the calculation of such estimates and predictors: *recursive methods* (associated with Markov models) and *generating function methods* (associated with cases in which structure is time-invariant and prediction errors are stationary).

If *x* and *y* are random vectors of zero mean then we shall use cov(*x, y*) to denote the crosscovariance matrix $E(xy')$ and shall write cov(*x, x*) simply as cov(*x*).

Recursive Methods: Markov Models and the Kalman Filter

Consider the dynamic equation, typical of many econometric models:

$$x_{t+1} = Ax_t + \in_{t+1} \qquad (1)$$

Here the process variable *x* is supposed to be a vector, and so *A* a corresponding square matrix, and ε_t is assumed to be vector white noise of zero mean and with covariance matrix *N*. One special feature of this model is that it is linear; another is that it is Markov (at least if ε is Gaussian). This is, that *x* is a state variable which constitutes a complete description, in that all aspects of the future which can be predicted from

$$X_t = (x_t, x_{t-1}, \ldots)$$

can also be predicted from x_t.

Suppose indeed that just $(x_t, x_{t-1}, \ldots)$ is observable at time *t*, so that $Y_t = X_t$. From (1) we deduce the solution

$$X_{t+m} = A^m x_t + \sum_{s=0}^{m-1} A^s \in_{t+m-s} \qquad (2)$$

for a future value $x_{1\ +\ m)}$ in terms of the current state x_t and future noise. Now, the white noise character of ε implies that $E[\varepsilon_{t+\tau}|X_t] = 0 \quad$ for $\tau > 0$. That is, a future noise value is unpredictable in that it has no better predictor than its unconditioned mean value: zero. We deduce then from (2) the simple expression for the predictor $x_{t+m}^{(t)} = E[x_{t+m}|X_t]$:

$$x_{t+m}^{(t)} = A^m x_t \quad (m \geqslant 0) \tag{3}$$

This obeys the recursion in *m:*

$$x_{t+m+1}^{(t)} = A x_{t+m}^{(t)} \quad (m \geqslant 0) \tag{4}$$

In other words, one predicts into the future just by solving the dynamic equation (1) with future noise set equal to its best prediction: zero. The predictor (3) would be *exact* for sequences $\{x_t\}$ generated from the noise-free version of (1):

$$x_{t+1} = A x_t \tag{5}$$

In the case (1) of noisy dynamics we see from (2) and (3) that the m-step prediction error will have covariance matrix

$$\mathrm{cov}\left[x_{t+m}^{(i)} - x_{t+m}\right] = \sum_{s=0}^{m-1} A^s N (A\prime)^s. \tag{6}$$

All these results remains true whatever the nature of A (and indeed have analogues if A is timevarying). If A has all its eigenvalues inside the unit circle, then system (1) is stable and will generate a stationary process. In other cases, $\{x_t\}$ will not be stationary, but conclusions (3), (4) and (6) still hold, and the prediction errors are stationary in time. For example, if A has a k-fold eigenvalue at unity, then the noise-free equation (5) will generate a polynomial trend in time of degree k – 1 (which (3) will predict exactly) and the actual dynamic equation (1) will generate a disturbed such trend. If A has other eigenvalues on the unit circle, then (1) will generate disturbed cyclicities (*undamped*, and possibly of polynomially increasing amplitude). Allowance of these possibilities provides the most ‘structural’ way of incorporating trends and seasonalities. If A has eigenvalues outside the unit circle, then (5) will have exponentially growing solutions (again predicted exactly by (3)) and (1) will have disturbed such solutions.

However, it will seldom be the case that the full state variable x will be observable. One must in general regard x as the state variable of an ideal Markov model, and indeed as a latent variable, which can be observed only partially. The usual and natural assumption is that, at time t, one can observe a vector y_t related to x by

$$y_{t+1} = C x_t + \eta_{t+1} \tag{7}$$

where $\{\varepsilon_1, \eta_t\}$ jointly constitute vector white noise with covariance matrix

$$\mathrm{cov}\begin{pmatrix} \in \\ \eta \end{pmatrix} = \begin{pmatrix} N & L \\ L' & M \end{pmatrix}. \tag{8}$$

Relations (1) and (7) together express process and observation structure.

Let us denote $x_t^{(t)}$ by $\widehat{x}_t$ the estimate of current state based on current observations. From (1), (7) and the properties of LLS estimates one can deduce that $\widehat{x}_t$ obeys the recursion

$$\widehat{x}_{t+1} = A\widehat{x}_t + H_t\left(y_{t+1} - C\widehat{x}_t\right) \tag{9}$$

where the matrix H_t is determined in terms of

$$V_t = \mathrm{cov}(\widehat{x}_t - x_t) \tag{10}$$

by the recursions

$$\begin{aligned} V_{t+1} = \quad & N + A V_t A' - (L + A V_t C') \\ & \times (M + C V_t C\prime)^{-1} (L' + C V_t A') \end{aligned} \tag{11}$$

$$H_t = (L + A V_t C')(M + C V_t C\prime)^{-1}. \tag{12}$$

Relation (9) constitutes the celebrated *Kalman filter* (a ‘filter’ being an operation for generating one sequence from another; in this case $\{\widehat{x}_t\}$ from $\{y_t\}$. The shortest of its many proofs is quite short (see, for example, Whittle 1983, p. 147).

There are many points to be made aboutthe Kalman filter. It is an updating relation, to be used in real time, to produce the new estimate of current state $\widehat{x}_{t+1}$ from the old one $\widehat{x}_t$ as a new observation $y_{t+ 1}$ becomes available. It takes the form of the dynamic equation (1) of the model itself, but driven, not by noise, but by the *innovation* $\left(y_{t+1} - C\widehat{x}_t\right)$. The innovation is to be interpreted as that part of the new observation $y_{t\,+\,1}$ which is not predictable from previous observations Y_t.

Once one has the state estimate, then prediction is simple. Because both process noise ε and observation noise η are supposed white one has

$$\begin{array}{rcll} x_{t+m}^{(t)} & = & A^m\widehat{x}_t, & (m \geqslant 0) \\ y_{t+m}^{(t)} & = & CA^{m-1}\widehat{x}_t, & (m > 0). \end{array} \tag{13}$$

Recursion (9) and relations (13) explain between them virtually all recursions (in t or m) between predictors to be found in the literature.

If appropriate conditions are satisfied (referred to as *observability* or *detectability* conditions), then, in the absence of plant and observation noise, the error in state estimate $\widehat{x}_t - x_t$ will tend to zero with increasing t. Under these same conditions the matrices V_t and H_t will tend to limit values V and H in the noisy case. Further, the matrix $\Omega = A - HC$ will be a stability matrix (so that Ω^s tends to zero exponentially fast with increasing s)

The Kalman filter (9) can then be written

$$\widehat{x}_{t+1} = \Omega\widehat{x}_{t+1} + Hy_{t+1} \tag{14}$$

with ‘solution’

$$\widehat{x}_t = \sum_{s=0}^{\infty} \Omega^S H y_{t-s}. \tag{15}$$

are of state and observation, and not of the parameters of the model itself, which are presumed known for present purposes.

$$g_{xy}(z) = \sum_{s=-\infty}^{\infty} z^s \mathrm{cov}(x_t, y_{t-s}),$$

a matrix function of the scalar marker variable z. Generating functions such as $g_{xy}(z)$ are closely related to Fourier ideas and the frequency concept; the power series becomes a Fourier series if we set $z = \exp(i\omega)$.

Suppose that $g_{xx}(z)$ has a *canonical factorization*

$$g_{xx}(z) = B(z)B\left(z^{-1}\right)' \tag{16}$$

where both $B(z) = \sum_{s=0}^{\infty} b_s z^s$ and $B(z)^{-1}$ are analytic in $|z| \leq 1$. Then the stationary process $\{x_t\}$ has both a *moving average representation*

$$x_t = \sum_{s=0}^{\infty} b_s \varepsilon_{t-s}, \tag{17}$$

where $\{\varepsilon_t\}$ is white noise with $\mathrm{cov}(\varepsilon) = I$ and an *autoregressive representation*

$$\sum_{s=0}^{\infty} a_s x_{t-s} = \in_t \tag{18}$$

where a_s is the coefficient of z^s in the expansion of $B(z)^{-1}$ in non-negative powers of z.

Suppose X_t is the observable at time t. Then, by the argument which led us from (2) to (3), the optimal predictor is:

$$x_{t+m}^{(t)} = \sum_{s=m}^{\infty} b_s \in_{t+m-s} \tag{19}$$

and this can be expressed explicitly in terms of X_t:

$$x_{t+m}^{(t)} = \sum_{s=m}^{\infty} \gamma_s x_{t-s} \tag{20}$$

by using (18) to express the ε variables of (19) in terms of the x variables. One can express this solution for the optimal prediction coefficients γ_s in generating function form

$$\sum_{s=m}^{\infty} \gamma_s z^s = \left[z^{-m}B(z)\right]_+ B(z)^{-1}, \tag{21}$$

where the operator $[]_+$ has the effect that all terms in negative powers of z in the series enclosed by the brackets are discarded.

Suppose, as is more usual, that x_t is not completely observable, but that at time t one has observed the values $y_t, y_{t-1} \ldots$ of some associated variable y. Then the predictor will now have the form $x_{t+m}^{(t)} = \sum_{s=0}^{\infty} \gamma_s y_{t-s}$, and the generalization of solution (21) is:

$$\sum_{s=m}^{\infty} \gamma_s z^s = \left[z^{-m} g_{xy}\left(g_{yy}^{-}\right)^{-1}\right]_{+}\left(g_{yy}^{+}\right)^{-1}. \qquad (22)$$

Here we have omitted the z-arguments for simplicity, and have assumed that the matrix generating function $g_{yy}(z)$ has canonical factorization

$$g_{yy} = g_{yy}^{+} g_{yy}^{-}. \qquad (23)$$

$$x_{t+m}^{(t+1)} = x_{t+m}^{(t)} + H\left(y_{t+1} - y_{t+1}^{(t)}\right) \qquad (24)$$

$$\begin{aligned} A(T)x_t &= \varepsilon_t \\ y_t + C(T)x_t &= \eta_t \end{aligned} \qquad (25)$$

where $A(T) = \sum_{s=0}^{\infty} A_s T^s, C(T) = \sum_{s=0}^{\infty} C_s T^s$, and T is the backwards translation operator, with effect $Tx_t = x_{t-1}$. We make the same assumptions about the noise variables as before: that they are white with covariance matrix (8).

Then appeal to the fact that $x^{(t)}$ can be regarded as an ML estimator (in the Gaussian case) as well as an LLS estimator leads to the conclusion that it satisfies a recursion

$$\begin{bmatrix} N & L & A(T) \\ L' & M & C(T) \\ A\left(T^{-1}\right)' & C\left(T^{-1}\right)' & 0 \end{bmatrix} \begin{bmatrix} \lambda \\ \mu \\ x \end{bmatrix}_{\tau}^{(t)} = \begin{bmatrix} 0 \\ y_\tau \\ 0 \end{bmatrix}, \quad (\tau \le t) \qquad (26)$$

(see Whittle 1983, p. 155). Here λ and μ specify Lagrangian multiplier sequences whose significance will emerge shortly, and the lag operator T operates on the running argument τ not on the fixed value t.

Note that relations (26) determine estimates of past and present stat $x_\tau^{(t)}(\tau \le t)$. However, once these have been derived, then predictors are easily calculated recursively from:

$$A(T)x_\tau^{(t)} = 0 \quad (\tau > t) \qquad (27)$$

Equivalently, (26) can be regarded as holding for all τ with λ, μ set equal to zero for $\tau > t$.

Relations (26) constitute an equation system to be solved, semi-infinite if observation indeed extends back into the indefinite past. Write the system as:

$$\Phi(T)\xi_\tau = \zeta_\tau \qquad (28)$$

Then a reduction that provides as explicit a solution for $x_t^{(t)}$ as is possible in the general case in the following. Suppose that the Hermitian matrix generating function $\Phi(z)$ has canonical factorization

$$\Phi(z) = \Phi^{+}(z)\Phi^{-}(z) \qquad (29)$$

Then under generalized observability conditions it is permisible to partially invert (26) to:

$$\Phi^{-}(T)\xi_\tau = \Phi^{+}(T)^{-1}\zeta_\tau \quad (\tau \le t) \qquad (30)$$

with the formal end condition $\xi_\tau = 0(\tau > t)$. Relation (30) for $\tau = t$ gives an explicit solution for $x_t^{(t)}$ in terms of Y_t; the relation for $\tau = t - 1$ then determines $x_{t-1}^{(t)}$ etc.

To see the wider significance of this approach one must consider the wider purpose of prediction. One will usually require predictions (or estimates of unobservables) to support actions. Suppose actions are chosen to minimize the expections $E(Q)$ of some quadratic function Q of process and action variables. If follows then from the certainty equivalence theorem that LLS (or ML) predictors are the optimal ones to use. However, suppose that one instead minimizes a criterion $-\theta^{-1}\log E\exp(-\frac{1}{2}\theta Q)$. Here θ is a measure of risk-sensitivity, implying risk-seeking behaviour (optimism) for $\theta > 0$ and risk-aversion (pessimism) for $\theta < 0$. This risk-sensitivity modifies the estimators, which we shall now term *minimal stress* estimates, for reasons explained in Whittle and Kuhn (1986). Remarkably, the above analysis still goes through, with the simple change that Φ has the modified definition

$$\Phi(T)=\begin{bmatrix} N & L & A(T)\\ L' & M & C(T)\\ A\left(T^{-1}\right)' & C\left(T^{-1}\right)' & -\theta R\end{bmatrix}.$$

Here R is a matrix corresponding to a component $\sum_t (x' Rx)_t$of Q which penalizes deviations of the process variable from a desired profile. Moreover, one can now establish the identity

$$\begin{bmatrix}N & L\\ L' & M\end{bmatrix}\begin{bmatrix}\lambda\\ \mu\end{bmatrix}_\tau^{(t)} = \begin{bmatrix}\varepsilon\\ \eta\end{bmatrix}_\tau^{(t)}$$

which relates λ and μ to the minimal stress estimates of noise.

See Also

- ▶ Forecasting
- ▶ Macroeconometric Models

Bibliography

Durbin, J. 1984. Present position and potential developments: Some personal views on time series analysis. *Journal of the Royal Statistical Society Series* 147: 161–173. This survey article contains a substantial reference list.

Whittle, P. 1983. *Prediction and regulation*, 2nd ed. Oxford: Blackwell and University of Minnesota Press.

Whittle, P., and J. Kuhn. 1986. A Hamiltonian formulation of risk-sensitive linear/quadratic/Gaussian control. *International Journal of Control* 43: 1–12.

Prediction Formulas

Charles H. Whiteman and Kurt F. Lewis

Abstract

Prediction formulas for multi-step forecasts and geometric distributed leads of stationary time series are derived using classical, frequency domain methods. Starting with the Wold representation, optimal squared-error loss predictions are derived using the analytic function theory approach of Whittle. This approach is easily adapted to the problem of making predictions that are robust under model misspecification. Forecasts and expected present value calculations are illustrated under both objectives for low-order autoregressive and moving average processes.

Keywords

Blaschke factors; Contour integral; Cross-equation restrictions; Distributed leads; Frequency domain problems; Least squares; Linear least squares projection; Minimum norm interpolation problem; Min-max problem; Misspecification; Prediction formulas; Rational expectations; Riesz–Fisher th; Robustness; Squared-error loss optimal prediction; Time domain problems; Wiener–Hopf equation; Wold decomposition th; Wold representation

JEL Classifications

C1

Introduction

This article reviews the derivation of formulas for linear least squares and robust prediction of stationary time series and geometrically discounted distributed leads of such series. The derivations employed are the classical, frequency-domain procedures employed by Whittle (1983) and Whiteman (1983), and result in nearly closed-form expressions. The formulas themselves are useful directly in forecasting, and have also found uses in economic modelling, primarily in macroeconomics. Indeed, Hansen and Sargent (1980) refer to the cross-equation restrictions connecting the time series representation of driving variables to the analogous representation for predicting the present value of such variables as the 'hallmark of rational expectations models'.

The Wold Representation

Suppose that $\{x_t\}$ is a covariance-stationary stochastic process and assume (without loss of generality) that $Ex_t = 0$. Covariance stationarity ensures that first and second unconditional moments of the process do not vary with time. Then, by the Wold decomposition theorem (see Sargent 1987, for an elementary exposition and proof), x_t can be represented by:

$$x_t = \sum_{j=0}^{\infty} a_j \varepsilon_{t-j} \tag{1}$$

with

$$a_0 = 1, \sum_{j=0}^{\infty} a_j^2 < \infty$$

and

$$\varepsilon_t = x_t - P(x_t | x_{t-1}, x_{t-2}, \ldots), E\varepsilon_t^2 = \sigma^2$$

where $P(x_t|x_{t-1}, x_{t-2}, \ldots)$ denotes the linear least squares projection (population regression) of x_t on x_{t-1}, x_{t-2} , . . . Here, 'represented by' need not mean 'generated by', but rather 'has the same variance and covariance stmcture as'. By construction, the 'fundamental' innovation ε_t is uncorrelated with information dated prior to t, including earlier values of the process itself: $E\varepsilon_t\varepsilon_{t-s} = 0 \; \forall \; s > 0$. This fact makes the Wold representation very convenient for computing predictions. The convolution in (1) is often written $x_t = A(L)\varepsilon_t$ using the polynomial $A(L) = \sum_{j=0}^{\infty} a_j L^j$ in the 'lag operator' L, where $L\varepsilon_t = \varepsilon_{t-1}$.

Squared-Error Loss Optimal Prediction

The optimal prediction problem under squared-error loss can be thought of as follows. Given $\{x_t\}$ with the Wold representation (1) we want to find the stochastic process y_t,

$$y_t = \sum_{j=0}^{\infty} c_j \varepsilon_{t-j} = C(L)\varepsilon_t$$

that will minimize the squared forecast error of the h-step ahead prediction

$$\min_{\{y_t\}} E(x_{t+h} - y_t)^2.$$

Equivalently, the problem can be written as

$$\min_{\{y_t\}} E\left(L^{-h} x_t - y_t\right)^2$$

or

$$\min_{\{c_j\}} E\left(L^{-h} \sum_{j=0}^{\infty} a_j \varepsilon_{t-j} - \sum_{j=0}^{\infty} c_j \varepsilon_{t-j}\right)^2. \tag{2}$$

The problem in (2) involves finding a *sequence* of coefficients in the Wold representation of the unknown prediction process y_t, and is referred to as the *time domain problem*. By virtue of the Riesz–Fisher theorem (see again Sargent 1987, for an exposition), the time-domain problem is equivalent to a *frequency domain problem* of finding an analytic function $C(z)$ on the unit disk $|z| \leq 1$ corresponding to the 'z-transform' of the $\{c_j\}$ sequence

$$C(z) = \sum_{j=0}^{\infty} c_j z^j$$

that solves

$$\min_{C(z) \in H^2} \frac{1}{2\pi i} \oint |z^{-h} A(z) - C(z)|^2 \frac{dz}{z} \tag{3}$$

where H^2 denotes the Hardy space of square-integrable analytic functions on the unit disk, and $\oint$ denotes (counterclockwise) integration about the unit circle. The requirement that $C(z) \in H^2$ ensures that the forecast is causal, and contains no future values of the ε's; this is equivalent to the requirement that $C(z)$ have a well-behaved power series expansion in non-negative powers of z.

Each formulation of the problem is useful, as often one or the other will be simpler to solve.

This stems from the fact that convolution in the time domain becomes multiplication in the frequency domain and vice versa. To see this, consider the two sequences $\{g_k\}_{k=-\infty}^{\infty}$ and $\{h_k\}_{k=-\infty}^{\infty}$. The convolution of $\{g_k\}$ and $\{h_k\}$ is the sequence $\{f_k\}$, in which a typical element would be:

$$f_k = \sum_{j=-\infty}^{\infty} g_j h_{k-j}.$$

The z-transform of the convolution is given by

$$\begin{aligned}
\sum_{k=-\infty}^{\infty} f_k z^k &= \sum_{k=-\infty}^{\infty}\left(\sum_{j=-\infty}^{\infty} g_j h_{k-j}\right) z^k \\
&= \sum_{k=-\infty}^{\infty}\sum_{j=-\infty}^{\infty} g_j z^j h_{k-j} z^{k-j} \\
&= \sum_{(k-j)=-\infty}^{\infty}\sum_{j=-\infty}^{\infty} g_j z^j h_{k-j} z^{k-j} \\
&= \sum_{s=-\infty}^{\infty}\sum_{j=-\infty}^{\infty} g_j z^j h_s z^s \,(\text{Substituting} \quad s = k-j) \\
&= \sum_{s=-\infty}^{\infty} h_s z^s \sum_{j=-\infty}^{\infty} g_j z^j = g(z)h(z).
\end{aligned}$$

Thus the 'z-transform' of the convolution of the sequences $\{g_k\}$ and $\{h_k\}$ is the product of the z-transforms of the two sequences.

Similarly, the z-transform of the product of two sequences is the convolution of the z-transforms:

$$\sum_{k=-\infty}^{\infty} g_k h_k z^k = \frac{1}{2\pi i}\oint g(p)h(z/p)\frac{dp}{p}.$$

To see why this is the case, note that

$$g(p)h(z/p)p^{-1} = \sum_{j=-\infty}^{\infty} g_j p^j \sum_{k=-\infty}^{\infty} h_k z^k p^{-k-1},$$

implying

$$\begin{aligned}
&\frac{1}{2\pi i}\oint g(p)h(z/p)p^{-1}dp \\
&= \frac{1}{2\pi i}\oint \sum_{j=-\infty}^{\infty}\sum_{k=-\infty}^{\infty} g_j h_k z^k p^{j-k-1} dp.
\end{aligned}$$

But all of the terms vanish except where $j = k$ because

$$\frac{1}{2\pi i}\oint z^k \frac{dz}{z} = 0$$

except when $k = 0$. To see why, let $z = e^{i\theta}$. As θ increases from 0 to 2π, z goes around the unit circle. So, since $dz = ie^{i\theta}d\theta$, we have that

$$\begin{aligned}
\frac{1}{2\pi i}\oint z^k \frac{dz}{z} &= \frac{i}{2\pi i}\oint e^{i\theta k} d\theta \\
&= \begin{cases} 1 & \text{if} \quad k = 0 \\ \dfrac{1}{2\pi}\dfrac{1}{ik} e^{i\theta k}|_0^{2\pi} = 0 & \text{otherwise.} \end{cases}
\end{aligned}$$

Thus,

$$\begin{aligned}
\frac{1}{2\pi i}\oint g(p)h(z/p)p^{-1}dp &= \sum_{j=-\infty}^{\infty} g_j h_j z^j \frac{1}{2\pi i}\oint \frac{dp}{p} \\
&= \sum_{j=-\infty}^{\infty} g_j h_j z^j
\end{aligned}$$

by Cauchy's Integral formula.

The frequency domain formulas can now be used to calculate moments quickly and conveniently. Consider Ex_t^2:

$$\begin{aligned}
Ex_t^2 = E(A(L)\varepsilon_t)^2 &= E\left(\sum_{j=0}^{\infty} A_j \varepsilon_{t-j}\right)^2 \\
&= \sigma_\varepsilon^2 \sum_{j=0}^{\infty} A_j^2. \qquad (4)
\end{aligned}$$

The result in Eq. (4) comes from the fact that $E\varepsilon_t\varepsilon_{t-s} = 0, \quad \forall s \neq 0.$ Using the product-convolution relation, we see that

$$\begin{aligned}
\sum^{j=0\infty} A_j^2 &= \sum_{j=0}^{\infty} A_j^2 z^j|_{z=1} \\
&= \frac{1}{2\pi i}\oint A(p)A(z/p)\frac{dp}{p}|_{z=1} \\
&= \frac{1}{2\pi i}\oint A(p)A\left(p^{-1}\right)\frac{dp}{p} \\
&= \frac{1}{2\pi i}\oint |A(z)|^2 \frac{dz}{z}. \qquad (5)
\end{aligned}$$

Returning to the prediction problem, the task is to choose c_0, c_1, c_2, . . . to

$$\min_{\{c_j\}} \frac{1}{2\pi i} \oint |z^{-h}A(z) - \sum_{j=0}^{\infty} c_j z^j|^2 \frac{dz}{z}. \qquad (6)$$

The first order conditions for the optimization in expression (7) are

$$\begin{aligned} 0 &= \frac{1}{2\pi i} \oint \{z^j[z^h A(z^{-1}) - C(z^{-1})] \\ &+ z^{-j}[z^{-h}A(z) - C(z)]\} \frac{dz}{z} \\ &= \frac{1}{2\pi i} \oint z^{-j}\left[z^{-h}A(z) - C(z)\right] \frac{dz}{z} \\ &- \frac{1}{2\pi i} \oint p^{-j}\left[p^{-h}A(p) - C(p)\right] \frac{dp}{p} \end{aligned} \qquad (7)$$

for $j = 0, 1, 2, \ldots$, where the second integral is the result of a change of variable $p = z^{-1}$ so that $dp = \; -z^{-1}dz$, resulting in

$$\frac{dp}{p} = z\left(-z^{-2}dz\right) = -\frac{dz}{z}.$$

The result is that in the second integral, the direction of the contour integration is clockwise. Multiplying by -1 and integrating counterclockwise, the second integral becomes identical to the first, and we can write the set of first-order conditions as

$$\begin{aligned} 0 &= \frac{1}{\pi i} \oint z^{-j}\left[z^{-h}A(z) - C(z)\right] \frac{dz}{z} j \\ &= 0, 1, 2, \ldots \end{aligned} \qquad (8)$$

Define $F(z)$ such that

$$F(z) = z^{-h}A(z) - C(z) = \sum_{j=-\infty}^{\infty} F_j z^j.$$

From Eq. (8), it must be the case that all coefficients on non-negative powers of z equal zero:

$$F_j = 0, \quad j = 0, 1, 2, \ldots .$$

Multiplying by z^j and summing over *all* $j = 0$, ± 1, ± 2, . . . , we obtain

$$F(z) = \sum_{-\infty}^{-1} \qquad (9)$$

where the term on the right-hand-side of (9) represents an unknown function in negative powers of z. Thus

$$z^{-h}A(z) - C(z) = \sum_{-\infty}^{-1} \, ,$$

which is an example of a ‘Wiener–Hopf’ equation. Now apply the (linear) ‘plussing’ operator, $[\cdot]_+$, which means ‘ignore negative powers of z’ The unknown function in negative powers of z is ‘annihilated’ by this operation, resulting in

$$\begin{aligned} C(z) &= \left[z^{-h}A(z)\right]_+ \\ &= \left[z^{-h}a_0 + z^{-h+1}a_1 + z^{-h+2}a_2 + \ldots\right]_+ \\ &= \left[z^0 a_h + z^1 a_{h+1} + z^2 a_{h+2} + \ldots\right] \\ &= \sum_{j=h}^{\infty} a_j z^{j-h} \\ &= z^{-h}A(z) - pr\left[z^{-h}A(z)\right] \end{aligned}$$

where $pr[z^{-h}A(z)]$ is the principal part of the Laurent expansion of $z^{-h}A(z)$ about $z = 0$. (The principal part of the Laurent expansion about $z = 0$ is the part involving negative powers of z.) This provides a very simple formula for computing forecasts.

AR(1) Example

Suppose that $x_t = ax_{t-1} + \varepsilon_t$. This means that $A(z) = 1/(1 - az)$. In this case:

$$\begin{aligned} C(z) &= \left[z^{-h}A(z)\right]_+ \\ &= \left[z^{-h}(1 + az + a^2z^2 + \ldots)\right]_+ \\ &= a^h(1 + az + a^2z^2 + \ldots) \\ &= \frac{a^h}{(1 - az)} \end{aligned}$$

and the least squares loss predictor of x_{t+h} using information dated t and earlier is

$$P_t^{LS} x_{t+h} = y_t = C(L)\varepsilon_t = C(L)A^{-1}(L)x_t = a^h x_t.$$

The forecast error is

$$x_{t+h} - a^h x_t = \varepsilon_{t+h} + a\varepsilon_{t+h-1} + \dots \quad + a^{h-1}\varepsilon_{t+1},$$

which is serially correlated (for $h \geq 2$), but not correlated with information dated t and earlier.

MA(1) Example

Supposed that $x_t = \varepsilon_t - \alpha\varepsilon_{t-1}$, meaning $A(z) = 1 - az$. Thus,

$$C(z) = \left[z^{-h}A(z)\right] = \left[z^{-h}(1 - \alpha z)\right] = \begin{cases} \alpha \text{ if } & h = 1, \\ 0 & \text{otherwise.} \end{cases}$$

So, the best one-step ahead predictor is

$$\alpha\varepsilon_t = \alpha\left(1 + \alpha L + \alpha^2 L^2 + \dots\right)x_t$$

and the best predictor for forecasts of horizon two or more is exactly zero. For two-step-ahead (and beyond) prediction, the forecast error is x_{t+h} itself, which is serially correlated but not correlated with information dated t and earlier.

Least Squares Prediction of Geometric Distributed Leads

A prediction problem that characterizes many models in economics involves the expectation of a discounted value. Perhaps the most common and widely studied example is the present value formula for stock prices. Abstracting from mean and trend, suppose the dividend process has a Wold representation given by

$$d_t = \sum_{j=0}^{\infty} q_j \varepsilon_{t-j} = q(L)\varepsilon_t \quad E(\varepsilon_t) = 0, \qquad E\left(\varepsilon_t^2\right) = 1. \tag{10}$$

Assuming that the constant discount factor is given by γ, we have the present value formula

$$p_t = E_t \sum_{j=0}^{\infty} \gamma^j d_{t+j} = E_t\left(\frac{q(L)}{1 - \gamma L^{-1}}\varepsilon_t\right) = E_t\left(p_t^*\right). \tag{11}$$

The least-squares minimization problem the predictor faces is to find a stochastic process p_t to minimize the expected squared prediction error $E\left(p_t - p_t^*\right)^2$. In terms of the information known at date t, the agent's task is to find a linear combination of current and past dividends, or, equivalently, of current and past dividend innovations ε_t, that is 'close' to p_t^*. Writing $p_t = f(L)\varepsilon_t$, the problem becomes one of finding the coefficients f_j in $f(L) = f_0 + f_1 L + f_2 L^2 + \dots$ to minimize $E\left(f(L)\varepsilon_t - p_t^*\right)^2$. Using the method described in the previous section, the problem has an equivalent, frequency-domain representation

$$\min_{f(z)\in H^2} \frac{1}{2\pi i}\oint \left| \frac{q(z)}{1 - \gamma z^{-1}} - f(z)\right|^2 \frac{dz}{z}. \tag{12}$$

The first-order conditions for choosing f_j are, after employing the same simplification used in (7),

$$-\frac{2}{2\pi i}\oint z^{-j}\left[\frac{q(z)}{1 - \gamma z^{-1}} - f(z)\right]\frac{dz}{z} = 0, \quad j = 0, 1, 2, \dots. \tag{13}$$

Now define

$$H(z) = \frac{q(z)}{1 - \gamma z^{-1}} - f(z)$$

so that (13) becomes

$$-\frac{2}{2\pi i}\oint z^{-j} H(z)\frac{dz}{z} = 0.$$

Then multiplying by z^j and summing over all $j = 0, \pm 1, \pm 2, \dots$ as above, we obtain

$$H(z) = \frac{q(z)}{1 - \gamma z^{-1}} - f(z) = \sum_{-\infty}^{-1},$$

the Wiener–Hopf equation for this problem. Applying the plussing operator to both sides yields

$$\left[\frac{q(z)}{1-\gamma z^{-1}}\right]_+ - [f(z)]_+ = 0$$

implying

$$f(z) = \left[\frac{q(z)}{1-\gamma z^{-1}}\right]_+ = \left[\frac{zq(z)}{z-\gamma}\right]_+$$

because *f*(*z*) is, by construction, one-sided in non-negative powers of *z*. As in the previous section,

$$[A(z)]_+ = A(z) - P(z)$$

where *P*(*z*) is the principal part of the Laurent series expansion of *A*(*z*). To determine the principal part of $[(z-\gamma)^{-1}zq(z)]$, note that *zq*(*z*) has a well-behaved power series expansion about $z=\gamma$, where 'well-behaved' means 'involving no negative powers of $(z-\gamma)$'. Thus $[(z-\gamma)^{-1}zq(z)]$ has a power series expansion about $z=\gamma$ involving a single term in $(z-\gamma)^{-1}$:

$$\left(\frac{zq(z)}{z-\gamma}\right) = \frac{b_{-1}}{z-\gamma} + b_0 + b_1(z-\gamma)^1 + b_2(z-\gamma)^2 + \ldots .$$

The principal part here is the part involving negative powers of $(z-\gamma)$: $\mathrm{b}_{-1}(z-\gamma)^{-1}$. To determine it, multiply both sides by $(z-\gamma)$ and evaluate what is left at $z=\gamma$ to find $b_{-1}=\gamma q(\gamma)$. Thus

$$f(z) = \left[\frac{q(z)}{1-\gamma z^{-1}}\right]_+ = \left[\frac{zq(z)}{z-\gamma}\right]_+ = \frac{zq(z)-\gamma q(\gamma)}{z-\gamma}. \qquad (14)$$

The 'cross-equation restrictions' of rational expectations refer to the connection between the serial correlation structure of the driving process (here dividends) and the serial correlation structure of the expected discounted value of the driving process (here prices). That is, when dividends are characterized by *q*(*z*), prices are characterized by *f*(*z*), and *f*(*z*) depends upon *q*(*z*) as depicted in (14).

To illustrate how the formula works, suppose detrended dividends are described by a first-order autoregression; that is, that $q(L)=(1-\rho L)^{-1}$. Then

$$p_t = f(L)\varepsilon_t = \frac{Lq(L)-\gamma q(\gamma)}{L-\gamma}\varepsilon_t = \left(\frac{1}{1-\rho\gamma}\right)d_t. \qquad (15)$$

It is instructive to note that, while the pricing formula (15) makes p_t the best least squares predictor of p_t^*, the prediction errors $p_t - p_t^*$ will not be serially uncorrelated. Indeed

$$\begin{aligned} p_t - p_t^* &= \gamma\left\{\frac{Lq(L)-\gamma q(\gamma)}{L-\gamma} - \frac{q(L)}{1-\gamma L^{-1}}\right\}\varepsilon_t \\ &= \frac{-\gamma^2 q(\gamma)}{L-\gamma}\varepsilon_t = -\gamma^2 q(\gamma)\frac{L^{-1}}{1-\gamma L^{-1}}\varepsilon_t \\ &= -\gamma^2 q(\gamma)\left\{\varepsilon_{t+1} + \gamma\varepsilon_{t+2} + \gamma^2\varepsilon_{t+3} + \ldots\right\}. \end{aligned}$$

Thus the prediction errors will be described by a highly persistent (γ is close to unity) first-order autoregression. But because this autoregression involves *future* ε_t's, the serial correlation structure of the errors cannot be exploited to improve the quality of the prediction of p_t^*. The reason is that the predictor 'knows' the *model* for price setting (the present value formula) and the dividend process; the best predictor $p_t = E_t p_t^*$ of p_t^* tolerates' the serial correlation because the (correct) model implies that it involves *future* ε_t's and therefore cannot be predicted. If one only had data on the errors (and did not know the model that generated them), they would appear (rightly) to be characterized by a first-order autoregression; fitting an AR(1) (that is, the best *linear* model) and using it to 'adjust' p_t by accounting for the serial correlation in the errors $p_t - p_t^*$ would decrease the quality of the estimate of p_t^*. The reason is the usual one that the Wold representation for $p_t - p_t^*$ is not the economic model of $p_t - p_t^*$, and (correct) models always beat Wold representations. This also serves as a reminder of circumstances under which one should be willing to tolerate serially correlated errors: when one knows the model that generated them, and the model implies that they are as small as they can be made.

Robust Optimal Prediction of Time Series

The squared-error loss function employed to this point is appropriate for situations in which the model (either the time series model or the economic model) is thought to be correct. But in many settings the forecaster or model builder may wish to guard against the possibility of misspecification. There are many ways to do this; an approach popular in the engineering literature and recently introduced into the economics literature by Hansen and Sargent (2007) involves behaving so as to minimize the maximum loss sustainable by using an approximating model when the truth may be something else. The ‘robust’ approach to this involves replacing the squared-error loss problem

$$\min_{\{C(z)\}} \frac{1}{2\pi i}\oint |z^{-h}A(z) - C(z)|^2 \frac{dz}{z}$$

with the ‘min-max’ problem

$$\min_{\{C(z)\}} \sup_{|z|=1} |z^{-h}A(z) - C(z)|^2,$$

so that minimizing the ‘average’ value on the unit circle has been replaced by minimizing the max. This problem can also be written

$$\min_{\{C(z)\}} \sup_{|z|=1} |A(z) - z^h C(z)|^2.$$

This is known as the ‘minimum norm interpolation problem’ and amounts to finding a function $\varphi(z)$ to

$$\min ||\varphi(z)||_\infty$$

subject to the restriction that the power series expansion of $\varphi(z)$ matches that of $A(z)$ for the first $h - 1$ powers of z. This means that the following must hold:

$$\sum_{j=0}^{h-1} \varphi_j z^j = \sum_{j=0}^{h-1} a_j z^j. \tag{16}$$

Theorem 1 *The minimizing $\varphi(z)$ function is such that $|\varphi(z)|^2$ is constant on $|z| = 1$. Moreover,*

$$\varphi(z) = M \prod_{j=1}^{h} \frac{z - \alpha_j}{1 - \overline{\alpha}_j z}$$

where M, α_1, α_2, ... ,α_n are chosen to ensure that (16) holds.

Proof: see Nehari (1957).

To see that $\varphi(z)$ must be of the indicated form, note that the ‘Blaschke factors’ in the product have unit modulus:

$$\frac{z-\alpha_j}{1-\alpha_j z}\left(\frac{z^{-1}-\overline{\alpha}_j}{1-\alpha_j z^{-1}}\right) = \left(\frac{z-\alpha_j}{1-\alpha_j z}\right)\left(z^{-1}z\right)$$
$$\left(\frac{z^{-1}-\overline{\alpha}_j}{1-\alpha_j z^{-1}}\right) = \left(\frac{1-\alpha_j z^{-1}}{1-\alpha_j z}\right)\left(\frac{1-\alpha_j z}{1-\alpha_j z^{-1}}\right) = 1,$$

so that $|\varphi(z)|^2 = M^2$.

In the general h-step-ahead prediction problem, we have that

$$\varphi(z) = M \prod_{j=1}^{h-1} \frac{z - \alpha_j}{1 - \overline{\alpha}_j z} = A(z) - z^h C(z),$$

meaning that

$$C(z) = \frac{1}{z^h}\left(A(z) - M \prod_{j=1}^{h-1} \frac{z - \alpha_j}{1 - \overline{\alpha}_j z}\right).$$

This is analogous to the solution in the least-squares case, but, instead of subtracting the principal part of $z^{-h}A(z)$, we subtract a different function from $z^{-h}A(z)$. Note also that because

$$M \prod_{j=1}^{h-1} \frac{z - \alpha_j}{1 - \overline{\alpha}_j z}$$

matches the power series expansion of $A(z)$ up to the power z^{h-1}, $C(z)$ is of the form

$$C(z) = c_0 + c_1 z + c_2 z^2 + \ldots$$

Finally, note that the forecast error is serially uncorrelated because $\varphi(z)$ is constant on $|z| = 1$.

Example. AR(1)

Let

$$A(z) = \frac{1}{1 - az}.$$

For $h = 1$, we see that $\varphi(z) = A(z) - zC(z)$ must be constant on $|z| = 1$, and that $\varphi(0) = A(0) = 1$. Thus, $\varphi(z) = M = 1$, so that

$$C(z) = \frac{A(z) - 1}{z} = \frac{az}{(1 - az)z} = \frac{a}{1 - az},$$

which implies that the robust one-step ahead forecast is

$$y_t^R = ax_t,$$

which coincides with the best least-squares forecast. This equivalence between the robust and least-squares one-step ahead forecasts is to be expected because the best one-step-ahead least-squares forecast also has serially uncorrelated errors. For $h = 2$, we have that

$$\varphi(z) = \frac{M(z - \alpha)}{1 - \overline{\alpha} z}$$

where (again) $\varphi(0) = 1$, but now we also see that $\varphi'(0) = a$. Thus,

$$\varphi(0) = 1 = -\alpha M \Rightarrow M = -\frac{1}{\alpha},$$

and furthermore

$$\begin{aligned}\varphi'(0) = a &= \frac{(1 - \overline{\alpha} z)M - M(z - \alpha)(-\overline{\alpha})}{(1 - \overline{\alpha} z)^2}|_{z=0} \\ &= M - M(\alpha\overline{\alpha}) = M(1 - \alpha\overline{\alpha}).\end{aligned}$$

Therefore, the solution will have the property that

$$\begin{aligned}a = -\frac{1}{\alpha}(1 - \alpha\overline{\alpha}) \quad - a\alpha = 1 - \alpha\overline{\alpha} \;\; 0 \\ = 1 + a\alpha - \alpha\overline{\alpha}.\end{aligned}$$

That is, the roots are reciprocal pairs. Notice that the discriminant is positive $(a^2\alpha^2 + 4\alpha\overline{\alpha} > 0)$, meaning that we will always have a real solution, and we choose $|\alpha| < 1$. Then, we have that

$$\begin{aligned}C(z) &= \frac{1}{z^2}\left[\frac{1}{1 - az} - \frac{M(z - \alpha)}{1 - \alpha z}\right] \\ &= \frac{1}{z^2}\frac{1 - \alpha z - (1 - az)\left(1 - \frac{1}{\alpha}z\right)}{(1 - az)(1 - \alpha z)} \\ &= \frac{1 - \alpha z - 1 + az + \frac{1}{\alpha}z - \frac{a}{\alpha}z^2}{z^2(1 - az)(1 - \alpha z)} \\ &= \frac{-\frac{a}{\alpha}}{(1 - az)(1 - \alpha z)}.\end{aligned}$$

So, the robust prediction is given by

$$P_t^R x_{t+2} = -\frac{a}{\alpha}\sum_{j=0}^{\infty} \alpha^j x_{t-j},$$

in contrast to the least-squares prediction

$$P_t^{LS} x_{t+2} = a^2 x_t.$$

Example. MA(1)

Suppose that the process follows an MA(1), $x_t = \varepsilon_t - \beta\varepsilon_{t-1}$, and therefore $A(z) = 1 - \beta z$. The analysis from the previous example still holds, and all of the following are true:

$$\varphi(z) = \frac{M(z - \alpha)}{1 - \overline{\alpha} z}$$

while

$$\varphi(0) = 1 = -\alpha M \;\Rightarrow\; M = -\alpha^{-1}$$

and

$$\varphi'(0) = -\beta = -\frac{1}{\alpha}(1 - \alpha\overline{\alpha}).$$

Therefore,

$$0 = 1 - \alpha\beta - \alpha\overline{\alpha},$$

meaning that, again, we have real roots which are reciprocal pairs and we can choose $|\alpha| < 1$. Of course, α will depend upon the value of β, and we write $\alpha(\beta)$. Thus

$$\begin{aligned} C(z) &= \frac{1}{z^2}\left[1 - \beta z - \frac{M(z - \alpha(\beta))}{(1 - \alpha(\beta)z)}\right] \\ &= \frac{1}{z^2}\left[\frac{(1 - \beta z)(1 - \alpha(\beta)z)) - M(z - \alpha(\beta))}{1 - \alpha(\beta)z}\right] \\ &= \frac{1}{z^2}\left[\frac{1 - \beta z - \alpha(\beta)z + \beta\alpha(\beta)z^2 - Mz + M\alpha(\beta)}{1 - \alpha(\beta)z}\right] \\ &= \frac{\beta\alpha(\beta)}{1 - \alpha(\beta)z}. \end{aligned}$$

Therefore, we have the robust prediction

$$\begin{aligned} P_t^R x_{t+2} &= \frac{\beta\alpha(\beta)}{1 - \alpha(\beta)L}\varepsilon_t \\ &= \frac{\beta\alpha(\beta)}{1 - \alpha(\beta)L}[x_t + \beta x_{t-1} + \beta x_{t-2} + \ldots], \end{aligned}$$

while the least-squares prediction is the standard

$$P_t^{LS} x_{t+2} = 0.$$

Robust Prediction of Geometric Distributed Leads

Following the excellent treatment in Kasa (2001), a robust present-value predictor fears that dividends may not be generated by the process in (10), and so, instead of choosing an $f(z)$ to minimize the average loss around the unit circle, chooses $f(z)$ to minimize the maximum loss:

$$\begin{aligned} &\min_{f(z)\in H^\infty} \sup_{|z|=1} \left|\frac{q(z)}{1 - \gamma z^{-1}} - f(z)\right|^2 \\ &\Leftrightarrow \min_{f(z)\in H^\infty} \sup_{|z|=1} \left|\frac{zq(z)}{z - \gamma} - f(z)\right|^2. \end{aligned}$$

Unlike in the least squares case (14), where $f(z)$ was restricted to the class H^2 of functions finitely square integrable on the unit circle, the restriction now is to the class of functions with finite maximum modulus on the unit circle, and the H^2 norm has been replaced by H^∞ norm.

To begin the solution process, note that there is considerable freedom in designing the minimizing function $f(z)$: it must be well-behaved (that is, must have a convergent power series in non-negative powers of z on the unit disk), but is otherwise unrestricted. Recalling the Laurent expansion

$$\frac{zq(z)}{z - \gamma} = \frac{b_{-1}}{z - \gamma} + b_0 + b_1(z - \gamma) + b_2(z - \gamma)^2 + \ldots,$$

while in the least squares case $f(z)$ was set to ‘cancel’ all the terms of this series except the first, here $f(z)$ will be set to do something else. Now define the Blaschke factor $B_\gamma(z) = (z - \gamma)/(1 - \gamma z)$ and note that, because of the unit modulus condition, the problem can be written

$$\min_{\{f(z)\}} \sup_{|z|=1} \left|\frac{zq(z)}{1 - \gamma z} - \frac{z - \gamma}{1 - \gamma z}f(z)\right|^2.$$

Defining

$$T(z) = \frac{zq(z)}{1 - \gamma z}$$

we have

$$\begin{aligned} &\min_{f\in H^\infty} \sup_{|z|=1} |T(z) - B_\gamma(z)f(z)| \\ &\Leftrightarrow \min_{f\in H^\infty} ||T(z) - B_\gamma(z)f(z)||_\infty. \end{aligned}$$

Define the function inside the $||$’s as

$$\varphi(z) = T(z) - B_\gamma(z)f(z)$$

and note that $\varphi(\gamma) = T(\gamma)$. Thus the problem of finding $f(z)$ reduces to the problem of finding the smallest $\varphi(z)$ satisfying $\varphi(\gamma) = T(\gamma)$:

$$\min_{\varphi \in H^\infty} ||\varphi(z)||_\infty \text{ s.t. } \varphi(\gamma) = T(\gamma)$$

Theorem 2 (Kasa 2001). *The solution to* (17) *is the constant function* $\varphi(z) = T(\gamma)$.

Proof. To see this, first note that the norm of a constant function is the modulus of the constant itself. This is written as

$$||\varphi(z)||_\infty = ||T(\gamma)||_\infty = |T(\gamma)|^2. \tag{17}$$

Next, suppose that there exists another function $\Psi(z) \in H^\infty$, with $\Psi(\gamma) = T(\gamma)$ and also

$$||\Psi(z)||_\infty < ||\varphi(z)||_\infty. \tag{18}$$

Recall the definition of the H^∞ norm, and using Eqs. (17) and (18):

$$||\Psi(z)||_\infty = \sup_{|z|=1} |\Psi(z)|^2 < |T(\gamma)|^2.$$

The maximum modulus theorem states that a function f which is analytic on the disk U achieves its maximum on the boundary of the disk. That is

$$\sup_{z \in U} |f(z)|^2 \leq \sup_{z \in \partial U} |f(z)|^2.$$

Therefore, we can see that

$$\sup_{|z|<1} |\Psi(z)|^2 \leq \sup_{|z|=1} |\Psi(z)|^2 < |T(\gamma)|^2.$$

However, one of the values on the interior of the unit disk is $z = \gamma$, which can be inserted into the far left-hand-side of Eq. (6) to get the result

$$|\Psi(\gamma)|^2 \leq \sup_{|z|} = 1|\Psi(z)|^2 < |T(\gamma)|^2 \Rightarrow |\Psi(\gamma)|^2 < |T(\gamma)|^2.$$

This contradicts the requirement that $\Psi(\gamma) = T(\gamma)$. Therefore, we have verified that there does not exist another function $\Psi(z) \in H^\infty$ such that $\Psi(\gamma) = T(\gamma)$ and $||\Psi(z)||_\infty < ||\varphi(z)||_\infty$. □

Given the form for $\varphi(z)$, the form for $f(z)$ follows. After some tedious algebra, we obtain

$$\begin{aligned} f(z) &= \frac{T(z) - \varphi(z)}{B_\gamma(z)} \\ &= \frac{zq(z) - \gamma q(\gamma)}{z - \gamma} + \frac{\gamma^2}{1 - \gamma^2} q(\gamma) \end{aligned}$$

which is the least squares solution plus a constant. Thus the robust cross-equation restrictions likewise differ from the least squares cross-equation restrictions. After the initial period, the impulse response function for the robust predictor is identical to that of the least squares predictor. In the initial period, the least squares impulse response is $q(\gamma)$, while the robust impulse response is larger: $q(\gamma)/(1 - \gamma^2)$.

Because γ is the discount factor, and therefore close to unity, the robust impulse response can be considerably larger than that of the least squares response. Relatedly, the volatility of prices in the robust case will be larger as well. For example, in the first-order autoregressive case studied above,

$$\begin{aligned} p_t &= f(L)\varepsilon_t \\ &= \frac{1}{1 - \rho\gamma} d_t + \frac{\gamma^2}{(1 - \gamma^2)(1 - \rho\gamma)} \varepsilon_t \end{aligned} \tag{19}$$

from which the variance can be calculated as

$$\begin{aligned} \sigma^2(p_t) = & \left(\frac{1}{1 - \rho\gamma}\right)^2 \sigma^2(d_t) \\ & + \frac{2\gamma^2 - \gamma^4}{(1 - \rho\gamma)^2 (1 - \gamma^2)^2}. \end{aligned}$$

When the discount factor is large and dividends are highly persistent, the variance of the robust present value prediction can be considerably larger than that of the least squares prediction (the first term on the right alone).

Finally, recall that the least-squares present-value predictor behaved in such a way as to minimize the *variance* of the error $p_t - p_t^*$. Here, robust prediction results in an error with Wold representation

$$p_t - p_t^* = \gamma\left\{\frac{Lq(L) - \gamma q(\gamma)}{L-\gamma} + \frac{\gamma^2}{1-\gamma^2}q(\gamma) - \frac{q(L)}{1-\gamma L^{-1}}\right\}\varepsilon_t$$

$$= -\frac{\gamma q(\gamma)}{1-\gamma^2}\left\{\frac{1-\gamma L}{L-\gamma}\right\}\varepsilon_t.$$

The term in braces has the form of a Blaschke factor. Applying such factors in the lag operator to a serially uncorrelated process like ε_t leaves a serially uncorrelated result; thus the robust present value predictor has behaved in such a way that the resulting errors are white noise. Of course this comes at a cost: to make the error serially uncorrelated, the robust predictor must tolerate an error variance that is larger than the least squares error variance by a factor of $a^2/(1-\gamma^2)$, which can be substantial when γ is close to unity.

See Also

- ▶ Forecasting
- ▶ Robust Control

Bibliography

Hansen, L.P., and T.J. Sargent. 1980. Formulating and estimating dynamic linear rational expectations models. *Journal of Economic Dynamics and Control* 2: 7–46.

Hansen, L.P., and T.J. Sargent. 2007. *Robustness*. Princeton: Princeton University Press.

Kasa, K. 2001. A robust Hansen–Sargent prediction formula. *Economic Letters* 71: 43–48.

Nehari, Z. 1957. On bounded bilinear forms. *Annals of Mathematics* 65(1): 153–162.

Sargent, T.J. 1987. *Macroeconomic theory*. New York: Academic Press.

Whiteman, C.H. 1983. *Linear rational expectations: A user ’s guide*. Minneapolis: University of Minnesota Press.

Whittle, P. 1983. *Prediction and regulation by linear least-square methods*. 2nd ed. Minneapolis: University of Minnesota Press.

Prediction Market Design

David C. Croson

Abstract

This article surveys (*a*) the challenges of transitioning the results from prediction-market experiments under laboratory conditions to outside-world conditions devoid of laboratory controls, (*b*) the abilities of current (as of this writing, in 2007) implementations of prediction markets to address these challenges, and (*c*) opportunities for research into future market designs which are robust to these challenges.

Keywords

Experimental methods in economics; Goodhart’s Law; Information aggregation; Iowa electronic markets; Laboratory markets; Moral hazard; Planning; Prediction market design; Prediction markets; Subjective probability

JEL Classifications

C9

prediction markets have made the jump from the laboratory to the field. In these markets, participants bid on Arrow securities, which pay one dollar in one state of the world and zero dollars in others. Since the pioneering work of Plott and Sunder (1982), experiments have generated results that are consistent with the idea that prices in controlled laboratory settings track predicted probabilities formed from the aggregated information of all participants (Hayek 1945). Emboldened by this general correspondence of theory and the general ability of laboratory experiments to test and validate these theories (see Sunder 1995; Plott 2000), prediction markets have escaped the laboratory. Markets offering opportunities to make real-money investments in predicting financial indices, political election results, entertainment awards, world events, and even the minutiae of sporting contests (now viewed with some suspicion as a highbrow form of gambling) have appeared in strength.

The history and general theory of such prediction markets is covered in prediction markets. This article combines (*a*) the challenges of transitioning the results from prediction-market experiments under laboratory conditions to outside-world conditions devoid of laboratory controls, (*b*) the abilities of current (as of this writing, in 2007) implementations of prediction markets to address these challenges, and (*c*) opportunities for research into future market designs which are robust to these challenges.

While there will necessarily be differences between strictly controlled laboratory studies and the real-world phenomena that they model, four particular cautions should be noted when attempting to extend the predictive abilities of laboratory-generated results to real-world prediction markets. The incidence of any of these four conditions will frustrate our ability to 'read' participants' collective estimation of probabilities from the equilibrium market prices for their associated securities: (*a*) extended duration of capital commitments; (*b*) differing levels of capital commitment across participants; (*c*) strategic objectives other than trading profits; and (*d*) influence by market participants over events on which the contracts are conditioned.

First, laboratory markets generally clear in short periods of time, so that participants face little or no opportunity costs from not investing their capital elsewhere for the duration of the experiment. In an outside prediction market, the lack of this quick-clearing condition means that prices will not generally sum to unity even when the alternatives are a complete partitioning of the possibility space. Participants' capital is tied up in their investment until its resolution, and the opportunity cost of tying up capital while awaiting resolution may be substantial. The common practice is for the exchange provider to capture the float by collecting deposits in time t dollars and paying in time $t + 1$ dollars at a 1:1 ratio, rather than a 1:$1 + r_f$ ratio; this distorts prices away from their associated probabilities. The problem is particularly acute when the time to expiry is relatively long over a high proportion of the contract life. An attempted resolution would be for the market organizer to pay the risk-free rate of return on all such committed capital and to allocate a fixed proportion $(1 - v)$ of the total amount collected (and earned) pro rata to the winners, with the result that the sum of the prices should converge towards $(1 - v)$; probabilities can be renormalized accordingly. In a first step towards addressing this problem, InTrade, a trade exchange market for social, political, and financial events, offered credit interest (at three per cent per annum) on all committed balances beyond a certain threshold account size, thereby reducing the time-value handicap faced by early investors in long-dated contracts.

Second, experimental market participants are allocated fixed amounts of capital, with any variation deliberate on the experimenters' part. When the capital is not uniform across participants, the resulting market prices may diverge from an unbiased estimate of participants' subjective probabilities, a situation hotly debated in the prediction-markets literature (Wolfers and Zitzewitz 2004; Manski 2006). The objection to interpreting prices as probabilities arises because the market price is both an input to decisions outside the market and the result of equilibration among these traders. Risk-neutral investors with subjective probabilities above (below) the current market price would wish to buy (sell) the contract, and thus the equilibrium price must balance the total capital of the players on the two sides of the current price. The price will reflect the subjective probability of the investor of the marginal dollar rather than the average subjective probabilities of the potentially large number of participants. With equal capital endowments, the side that prefers the low-probability (inexpensive) side of the contract demands more contract quantity, driving the price down; with unequal capital endowments, this problem can be either masked or exacerbated. The fact that 'heavy hitters' with bigger budgets (or optimistic beliefs about long-shot events) get more 'votes' in this system obscures the information gathered from the other participants. At best, market prices in these situations can be interpreted as the dollar-weighted averages of participants' beliefs, rather than the simple median.

Consider, then, the task of extracting information from such a market. A prospective decision-maker does not observe the budget weights and will thus be unable to correct for them in the weighted average of collective opinion. A simple solution would make public the holdings of investors with concentrated positions of more than five per cent of outstanding contracts (similar to the 13-D filings required in US securities markets to indicate heavy individual or institutional ownership), or to report measures of concentration in contract ownership or short position. The Iowa Electronic Markets (IEM) has, from its inception, imposed strict capital-inflow restrictions through limiting account funding to 500 dollars (Berg et al. 2006) allowing the IEM to avoid this problem. Substantially larger sums (often in the thousands, if not tens of thousands, of dollars) can be instantly committed at other markets to back a financial investor's probability estimates (for example, those based on the Dow Jones Industrial Average at InTrade).

The third issue is the possibility of strategic manipulation of market prices to achieve an outside goal not shared by all market participants. In a laboratory, the incentives (monetary and otherwise) may be substantial or tiny, but they are by design completely separable from any participant-specific objectives in the outside world. In the absence of such separability, agents with objectives other than capital gain in the prediction market may participate strategically, conveying distorted information to those who rely on unbiased market prices. In a political campaign, for example, it may be worth substantially more to candidates to generate the impression of public support for their preferred campaign than to make a profit on their investments in the information market. Paradoxically, the more credence is given by the general public to the prices-as-probabilities predictions, the higher the incentives for strategic investors to distort price by devoting relatively modest amounts of their private capital to moving the market. (This is an application of Goodhart's Law, wherein indirect measures targeted as policy goals lose their predictive ability.) In extreme cases, this strategic investment may affect voters' decisions on participation and on candidate selection, thus becoming a self-fulfilling prophecy. This could not only achieve election goals, but also generate capital gains from the information-distorting investment. Therefore, the cost of such a manipulation campaign can be zero or even negative, increasing its attractiveness but sapping the market's predictive power. The design of a prediction market to aggregate opinions and subjective probabilities without encouraging such strategic behaviour remains an open problem.

The fourth issue is the possibility of hidden control of seemingly random events. Under experimental conditions, experimenters control many basic aspects of the study, including randomization of events that are supposed to be random. Since real-world prediction markets generally lack such controls, we must thus be especially cautious in interpreting prices-as-probabilities when certain individuals can profit by exploiting their disproportionate ability to influence the occurrence of the event that is being predicted. As Croson and Kunreuther (2000) note in the analogous situation in the insurance market for natural catastrophic disasters against those caused by terrorism or war, moral hazard can destroy the risk-hedging functions of these markets. The social desirability of such a prediction market subject to moral hazard depends crucially on whether the efficiency value of early warning (caused by the propagation, through price changes, of the inside information) outweighs the equity or efficiency costs resulting from perverse incentives.

Several Fortune 500 companies (for example, HP and Google) have recently implemented prediction markets within the firm. These markets are designed to aggregate information among many employees and thereby produce reasonably accurate estimates that would otherwise be difficult (or impossible) for any single decision-maker to form. Such attempts effectively illustrate Hayek's famous argument (1945) that central planning cannot replicate the effects of distributed information; one can hardly fault these firms' desire for an unbiased, distributed information-gathering mechanism that communicates a clear message to decision-makers. Corporate motivations for these markets seem primarily aimed at the noble goals of selecting among several alternative investments, informing investment decisions in complementary activities, or attempting to predict future competitive opportunities; they conspicuously and simultaneously risk deviating from controlled experiments, however, in *all four* of the dimensions offered warningly above. Until such markets can be designed to resolve quickly, enforce equal participation among organization members at different ranks or economic stations, disentangle 'in-market' gains from 'out-of-market' gains, and disallow participation by employees who can influence the outcome of the events on which contracts are conditioned, the equilibrium prices shown by these markets will be suspect as a measure of participants' subjective probabilities. Accordingly, unless we develop tools to separate participants' choices from their jobs (effectively creating a 'virtual laboratory' inside the firm) or to correct for the biases induced by these deviations (extracting accurate and useful decision-support information from an unavoidably distorted market), the aggregate value of corporate use of these admittedly promising tools in non-laboratory conditions will be limited, and corporate successes using this powerful technique will be determined as much by chance as by economic science.

In strictly controlled laboratory studies, these four divisive effects can be minimized: such studies are completed over short periods of time (making the discounting problem minuscule); participants can be allocated exogenously fixed amounts of capital; the payoffs from successful experimental investments can be separated from outside gains, and the incidence of random events can be kept unpredictable. As prediction markets gain wider acceptance, more active participants, and public credence in the world outside the laboratory, however, the ability to interpret prices as the participants' subjective probabilities of Arrow events becomes increasingly tenuous. To profit from prediction markets outside the laboratory, corporations and investors must combine their knowledge of the established economics of such markets with skills in evaluating investor psychology, probability models more accurate than those of rival participants, and methods of extracting information from noisy and potentially biased signals.

See Also

- ▶ Experimental Economics
- ▶ Experimental Methods in Economics
- ▶ Information Aggregation and Prices
- ▶ Moral Hazard
- ▶ Prediction Markets

Bibliography

Arrow, K.J. 1964. The role of securities in the optimal allocation of risk-bearing. *Review of Economic Studies* 31: 91–96.

Berg, J., R. Forsythe, F. Nelson, and T. Rietz. 2006. Results from a dozen years of election futures markets research. In *Handbook of experimental economic results*, ed. C. Plott and V. Smith. Amsterdam: North-Holland.

Croson, D., and H. Kunreuther. 2000. Customizing indemnity contracts and indexed cat bonds for natural hazard risks. *Journal of Risk Finance* 1: 24–41.

Hayek, F. 1945. The use of knowledge in society. *American Economic Review* 35: 519–530.

InTrade. Online. Available at http://www.intrade.com. Accessed 22 June 2007.

Iowa Electronic Markets. Online. Available at http://www.biz.uiowa.edu/iem/. Accessed 23 June 2007.

Manski, C.F. 2006. Interpreting the predictions of prediction markets. *Economic Letters* 91: 425–429.

Plott, C. 2000. Markets as information gathering tools. *Southern Economic Journal* 67: 1–15.

Plott, C., and V. Smith. 2006. *Handbook of experimental economic results*. Amsterdam: North-Holland.

Plott, C., and S. Sunder. 1982. Efficiency of experimental security markets with insider information: An application of rational expectations models. *Journal of Political Economy* 90: 663–698.

Roth, A., and J. Kagel. 1995. *Handbook of experimental economics*. Princeton: Princeton University Press.

Sunder, S. 1995. Experimental asset markets: A survey. In *Handbook of experimental economics*, ed. A. Roth and J. Kagel. Princeton: Princeton University Press.

Wolfers, J., and E. Zitzewitz. 2004. Prediction markets. *Journal of Economic Perspectives* 18(2): 107–126.

Prediction Markets

Justin Wolfers and Eric Zitzewitz

Abstract

Prediction markets, sometimes referred to as 'information markets', 'idea futures' or 'event futures', are markets where participants trade contracts whose payoffs are tied to a future event, thereby yielding prices that can be interpreted as market-aggregated forecasts. This article summarizes the recent literature on prediction markets, highlighting both theoretical contributions that emphasize the possibility that these markets efficiently aggregate dispersed information, and the lessons from empirical applications which show that market-generated forecasts typically outperform most moderately sophisticated benchmarks. Along the way, we highlight areas ripe for future research.

Keywords

Constant absolute risk aversion (CARA); Decision markets; Efficient market hypothesis; Favourite-longshot bias; Forecasting; Gallup Poll; Information aggregation; Iowa Electronic Market; Prediction; Prediction markets; Probability; Spread betting

JEL Classifications

D7; D8

Prediction markets, sometimes referred to as 'information markets,' 'idea futures' or 'event futures', are markets where participants trade contracts whose payoffs are tied to a future event, thereby yielding prices that can be interpreted as market-aggregated forecasts. For instance, in the Iowa Electronic Market traders buy and sell contracts that pay one dollar if a given candidate wins the election. If a prediction market is efficient, then the prices of these contracts perfectly aggregate dispersed information about the probability of each candidate being elected. Markets designed specifically around this information aggregation and revelation motive are our focus in this article.

Types of Prediction Market

The most famous prediction markets are the election forecasting markets run by the University of Iowa (Berg et al. 2006). Election forecasting provides a useful way to introduce a variety of different contract types, and Table 1, adapted from Wolfers and Zitzewitz (2004a), shows how different contracts can be designed to reveal various types of forecasts.

The three main types of contract link payoffs to the occurrence of a specific event (the incumbent wins the election), to a continuous variable (the vote share of the incumbent), or to a combination of the two, such as in spread betting (the vote share of the incumbent exceeds x per cent). In each case, the relevant contract will reveal the market's expectation of a specific parameter: a probability, a mean or a median, respectively. More complex contract designs can also be used to elicit alternative parameters. For instance, a family of winner-take-all contracts – each linked to different states of nature – can reveal the full probability distribution.

Prediction markets have been used to forecast elections, movie revenues, corporate sales, project completion, economic indicators and Saddam Hussein's demise. New corporate applications have emerged as firms have looked to markets to predict research and development outcomes, the success of new products, and regulatory outcomes. In the US public sector, the Pentagon attempted to use markets designed to predict geopolitical risks, although negative publicity stopped the project (Hanson 2006). An intriguing attempt to apply prediction markets to forecasting influenza outbreaks is detailed in Nelson et al. (2006). Rhode and Strumpf (2004) have detailed the existence of large-scale election betting as far back as the election of President Grant in 1868.

Prediction market contracts have been traded in a variety of market designs, including

Prediction Markets, Table 1 Contract types: estimating uncertain quantities or probabilities

Contract	Details	Example	Reveals market expectation of. . .	More general application
Binary option	Contract costs $$p$ Pays $1 if and only if event x occurs.	Event x: George Bush wins the popular vote.	Probability that event y occurs, $p\ (x)$.	Defining many events, $x_1, x_2, \ldots, x_n$ reveals probability distribution $F(x)$.
Index futures	Contract pays $$x$.	Contract pays $1 for every percentage point of the popular vote won by George Bush.	Mean value of outcome x: E[x].	Contract pays some function of x: $$g(x)$. Reveals specific moments, E[$g(x)$].
Spread betting	Contract costs $1 Pays $2 if $x > x^*$ Pays $0 otherwise. Bid according to the value of x^*.	Contract pays even money if Bush wins more than x^* % of the popular vote.	Median value of outcome, x.	$1 contract pays $$(1/q)$ if $x > x^*$w. Reveals specific quantile, $F_{1-q}(x)$.

continuous double auctions (both with and without market-makers), pari-mutuel pools, and bookmaker-mediated betting markets, or implemented as market-scoring rules.

Prediction Markets in Theory: Information Aggregation

The claim that prediction markets can efficiently aggregate information is based on the efficient market hypothesis. In certain cases, existing theoretical results regarding efficient capital markets can be applied directly. Grossman (1976) documents a set of sufficient conditions for the equilibrium price of index futures to summarize private information perfectly: in a market where traders with constant absolute risk aversion (CARA) utility functions each receive independent draws from a normal distribution about the true value of the asset, the market price fully summarizes their information.

Manski (2004) notes that much of the analysis of the price of binary options simply assumes that these revealed a market-based probability estimate, but that appropriate theoretical results are lacking. He illustrates the importance of this issue by way of an example where prediction market prices fail to aggregate information appropriately. In his model all traders are willing to risk exactly $100. Thus, if a contract paying $1 if an event occurs is selling for $0.667, then buyers each purchase 150 contracts, while sellers can afford to sell 300 contracts (at a price of $0.333). This can be an equilibrium only if there are twice as many buyers as sellers, implying that the market price must fall at the 33rd percentile of the belief distribution, rather than the mean. The same logic suggests that a prediction market price of π implies that $1 - \pi$ per cent of the population believes that the event has less than a π per cent chance of occurring. Clearly, the driving force in this example is the assumption that all traders are willing to risk a fixed amount.

Wolfers and Zitzewitz (2005a) provide sufficient conditions under which prediction market prices coincide with average beliefs among traders (and hence aggregate all information in the Grossman set-up). They consider individuals with log utility and initial wealth, y, who must choose how many prediction market securities, x, to purchase at a price, π, given that they believe that the probability of winning their bet is q:

$$Max\ EU_{j_{\{x\}}} = q_j Log\left[y + x_j(1-\pi)\right] + (1-q_j)Log\left[y - x_j\pi\right]$$
$$yielding: \quad x_j^* = y\frac{q_j - \pi}{\pi(1-\pi)}$$

The prediction market is in equilibrium when supply equals demand:

$$\int_{-\infty}^{\pi} y\frac{q-\pi}{\pi(1-\pi)}f(q)dq = \int_{\pi}^{\infty} y\frac{\pi-q}{\pi(1-\pi)}f(q)dq$$

If beliefs (q) and wealth (y) are independent, then this implies:

$$\pi = \int_{-\infty}^{\infty} qf(q)dq = \overline{q}.$$

Thus, under log utility, the prediction market price equals the mean belief among traders. If wealth is correlated with beliefs, then the prediction market price is equal to a wealth-weighted average belief. This finding is general in the sense that no assumptions are required about the distribution of beliefs, but it is also quite specific in that it holds only under log utility. Experimenting with a range of alternative utility functions and distributions of beliefs typically yields prediction market prices that diverge from the mean of beliefs by only a small amount.

Both the Manski and the Wolfers–Zitzewitz models are silent as to the sources of the different beliefs across traders, which allows them to sidestep the theoretical difficulty posed by Milgrom and Stokey (1982), namely, that under common beliefs no trade will occur. The logic of the 'no trade theorem' is simply that traders should always be wary that anyone seeking to trade with them possesses an information advantage, and hence should moderate their beliefs accordingly. Why there should be any trade in prediction markets remains an important open theoretical question. Wolfers and Zitzewitz (2006) provide a simple adaptation of the Kyle (1985) model in which trade is driven by uninformed outsiders with either hedging- or entertainment-driven demand for the prediction security, or by manipulators attempting to influence market prices.

Another important role of prediction markets is that potential trading profits provide an incentive for *information discovery*. Grossman and Stiglitz (1976) consider the case where information is expensive to garner. They point to the impossibility of prices being fully efficient: if prices fully reflect information, then there is no incentive for any trader to gather that information. Instead, they construct a model in which prices never fully reflect all of the information possessed by informed traders; in equilibrium the inefficiency in pricing is just sufficient to induce a proportion of traders to become informed.

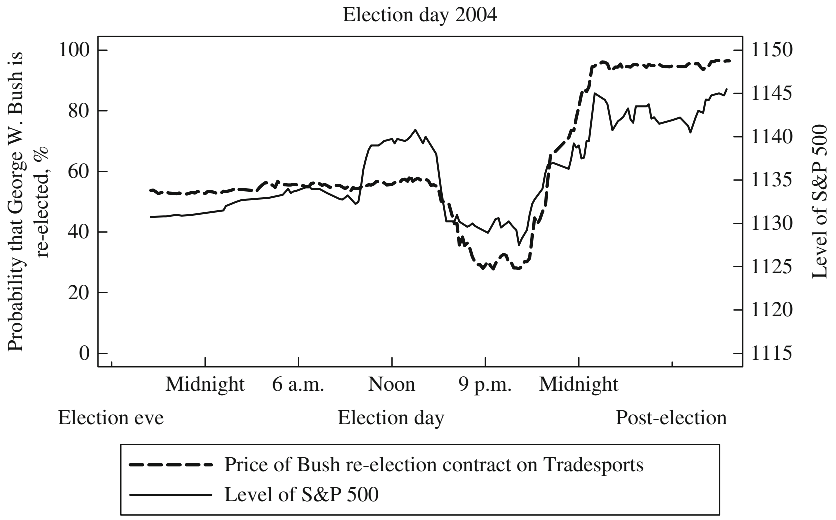
Another key advantage of prediction markets over alternative approaches to information aggregation is that they provide incentives for *truthful revelation* of beliefs. If prediction markets are to be used as inputs into future decisions, this may provide a countervailing incentive to trade dishonestly to manipulate prices. While such manipulation would typically lead the manipulator to lose money, Hanson and Oprea (2005) have shown that these losses increase the rewards for informed trading, which may ultimately increase the accuracy of prediction market prices.

Prediction Markets in Practice

While we are still accumulating evidence on the behaviour of prediction markets in different contexts, already a few generalizations can be drawn from existing, albeit piecemeal, evidence.

First, market prices tend to respond rapidly to new information. Figure 1 draws an interesting example from Snowberg et al. (2006): movements in the price of the Tradesports contract on the re-election of US President Bush, around election day, 2004. Early exit polls suggesting victory by John Kerry, the Democrat candidate, were leaked at around 3 p.m., and prices started to move immediately. Indeed, the figure shows that they moved in lockstep with prices on the much larger equity markets. As the count proceeded, it became clear that these early polling numbers were wrong, and the market reversed course sharply. This is only a single anecdote but is representative of the rapid incorporation of new information by prediction markets observed in many domains.

Second, in most cases, the time series of prices in these markets appears to follow a random walk, and simple betting strategies based on publicly available information appear to yield no profit opportunities. That is, these markets appear to meet the standard definition of weak-form efficiency.



Prediction Markets, Fig. 1 Bush's re-election prospects and the stockmarket (*Source*: Snowberg et al. (2006))

Third, the law of one price appears to (roughly) hold, and the few arbitrage opportunities that arise in these markets are fleeting and involve only small potential profits.

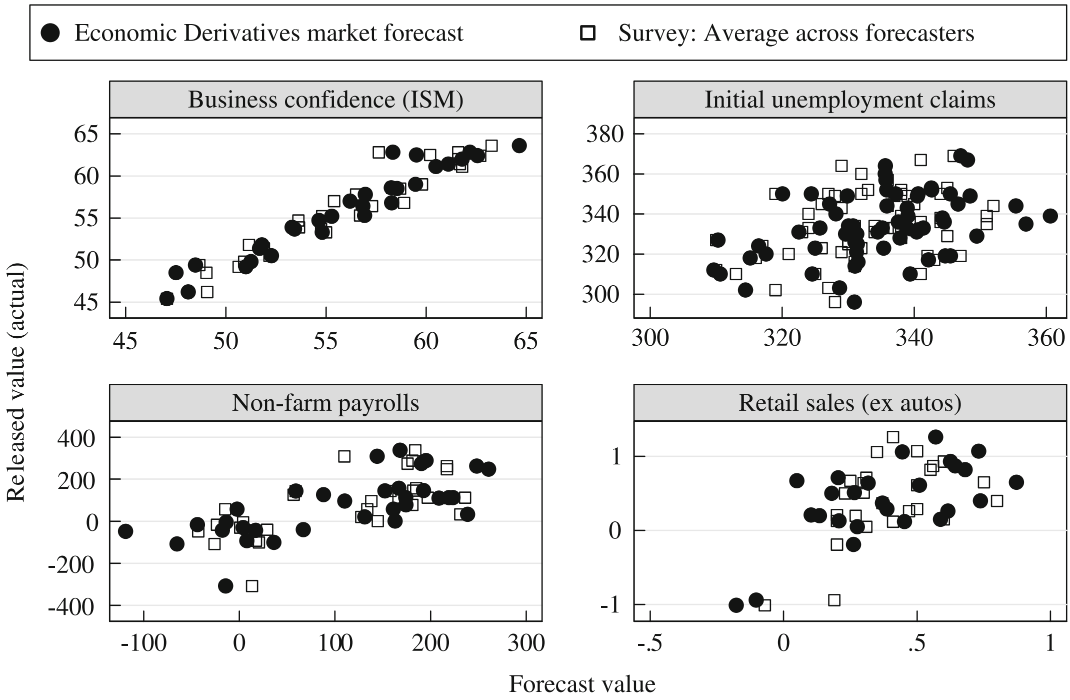
Fourth, attempts to manipulate these markets typically fail. Camerer (1998) attempted to manipulate pari-mutuel betting on horse races by canceling $500 or $1,000 bets at the last moment. Rhode and Strumpf (2006) report attempts by specific political campaigns to manipulate the election betting odds on their candidates in the large-scale betting markets operating in the early 20th century. They also analyse an attempt to manipulate the price of a Kerry victory on Tradesports in 2004, as well as their own attempts to manipulate prices on the Iowa Electronic Markets in 2000. Hanson et al. (2006) created experimental prediction markets in which several traders were given an incentive to raise the price. None of these attempts at manipulation had a discernible effect on prices, except during a short transition phase.

Finally, prediction markets usually provide quite accurate forecasts and have typically outperformed alternative prediction tools.

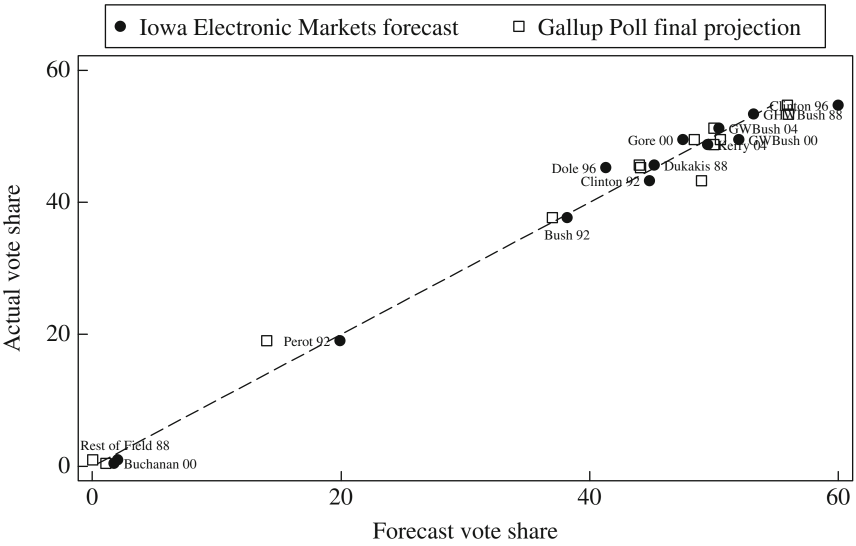
Figure 2 shows evidence collected by Gürkaynak and Wolfers (2005) on the relative performance of a prediction market (the 'Economic Derivatives' market established by Goldman Sachs and Deutsche Bank) and a survey of economists in predicting economic outcomes. They show that the market-based forecast encompasses the information in the survey-based forecasts. Moreover, the behavioural anomalies that have been noted in survey-based forecasts are not evident in the market-based forecasts.

Figure 3 compares the forecasting performance of the Iowa Electronic Markets and the Gallup Poll in predicting the outcomes of presidential elections in the United States. Over the 13 major candidacies from 1988 to 2004, the average absolute error of the market-based forecasts was 1.6 percentage points, while the corresponding number for the Gallup Poll was 1.9 percentage points. As Berg et al. (2003) discuss, the forecasting advantage of markets over the polls is probably even larger over long horizons, as polling numbers tend to be excessively volatile through the electoral cycle. The initial success of these forecasting methods in the United States has led to similar analysis of election forecasting markets in Austria, Australia, Canada, Germany, the Netherlands and Taiwan.

Tests of prediction markets and expert opinions have also been conducted in a range of other domains. The Hollywood Stock Exchange has generated forecasts of box-office success and of

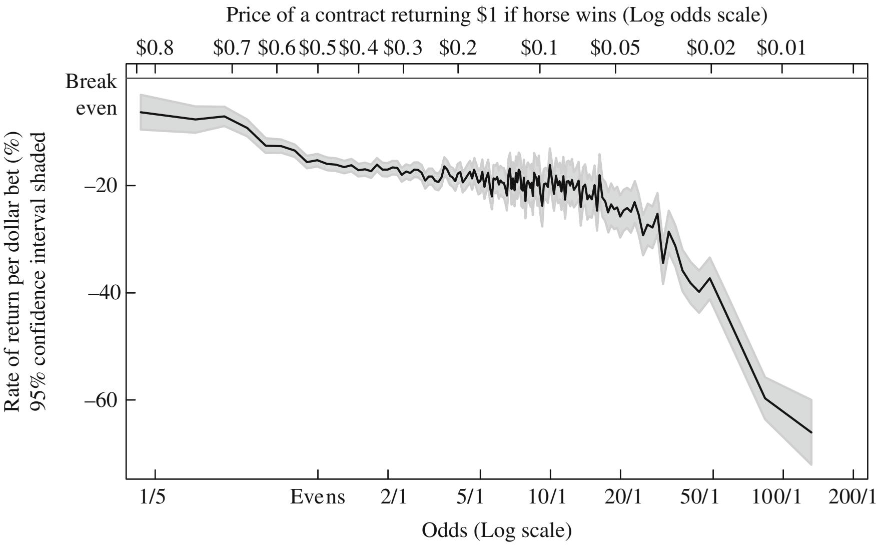


Prediction Markets, Fig. 2 Forecasting economic outcomes. Graphs by economic data series (*Source*: Gürkaynak and Wolfers (2005))



Prediction Markets, Fig. 3 Forecasting presidential elections. *Note*: Market forecast is closing price on election eve; Gallup forecast is final pre-election projection

Oscar winners that have been more accurate than expert opinions (Pennock et al. 2001). Both real and play-money markets have generated more accurate forecasts of the likely winners of NFL football games than all but a handful among 2,000 self-professed experts (Servan-Schreiber et al.



Prediction Markets, Fig. 4 Favourite-longshot bias: rate of return at different odds (*Source*: Trackmaster, Inc. Sample is all horse races in the United States, 1992–2002, n = 5,067,832 starts in 611,807 races)

2004). In the corporate context, the market established by Chen and Plott (2002) within Hewlett-Packard yielded more accurate sales forecasts than the firm's internal experts. Similarly, Ortner (1998) reports that an internal market correctly predicted that the firm would definitely fail to deliver on a software project on time, even when traditional planning tools suggested that the deadline could be met.

Despite this impressive evidence, there remain a number of documented pathologies in prediction markets. Figure 4 shows evidence from Snowberg and Wolfers (2005) of the 'favourite-longshot bias', which describes a tendency to overprice low-probability events. A similar tendency has been documented in a range of other market contexts, suggesting that some caution is in order in interpreting the prices of low probability events.

Laboratory experiments also find that, while prediction markets can be successful in some contexts (Plott and Sunder 1982), in others they may fail to aggregate information (Plott and Sunder 1988). Sunder (1995) and Plott (2000) provide excellent reviews of experimental prediction markets, including experiments showing market designs that lead to the appearance of bubbles, false equilibria or excess volatility.

Economic Analysis of Prediction Market Prices

Prediction markets are a useful way to elicit predictions, but how might they be used? The most direct form of inference involves simply using these predictions directly. For instance, forecasts of election outcomes may be of intrinsic interest.

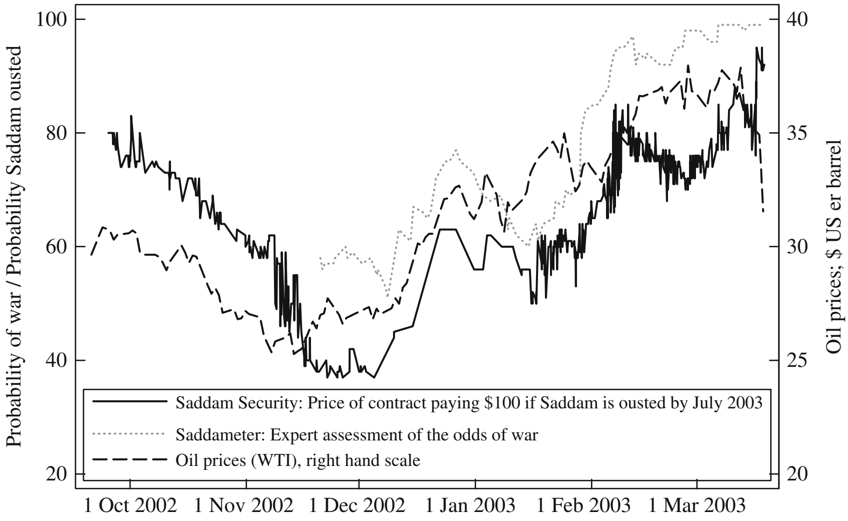
Some analyses have tried to link the time series of expectations elicited in prediction markets with time series of other variables, so as to isolate a causal influence. For instance, Roberts (1990) analyses changes in the betting odds posted by Ladbrokes on US President Ronald Reagan's re-election in 1984 and the returns to holding stocks in defence firms, inferring that Reagan led to more robust defence spending. Likewise, Herron et al. (1999) and Knight (2006) analyse the correlation of industry stock indices and individual stocks with movements in the 1992 and 2000 Iowa Electronic Markets US presidential election markets. Snowberg et al. (2006) conduct a similar analysis for the aggregate equity and bond markets at an intraday frequency, using the data shown in Fig. 1, to infer partisan impacts of the 2004 election. Slemrod and Greimel (1999) examine the effect on municipal bond prices of changes in the probability of a 1996 Republican nomination for Steve Forbes, whose 'flat tax' would have eliminated the tax exemption for municipal bond interest.

To move beyond *ex post* studies of elections, Wolfers and Zitzewitz (2005b) report on an *ex ante* analysis of the co-movement of oil and equity prices with a contract tracking the probability of a US attack on Iraq in 2002–3 (Fig. 5). The results suggest that a substantial war premium was built into oil prices (and a discount built into equities).

The contracts we have described thus far have depended on only one outcome. The same principles can be applied to contracts tied to the outcomes of more than one event. These contingent contracts potentially provide insight into the correlation between events. For instance, Wolfers and Zitzewitz (2004b) ran experimental markets on the online betting exchange Tradesports.com in the run-up to the 2004 US presidential election. In one example, they ran markets linked to whether George W. Bush would be re-elected, whether Al-Qaeda leader Osama bin Laden would be captured prior to the election, and whether *both* events would occur. These markets suggested a 91 per cent chance of Bush being re-elected *if* Osama had been found, but a 67 per cent unconditional probability. Berg and Reitz (2003) report on contracts whose payoff was linked to 1996 Democratic vote shares conditional on different potential Republican nominees; on the basis of these prices they argue that alternative nominees, such as Colin Powell, would have outperformed Bob Dole, the actual nominee.

The potential to apply these markets to determine the consequences of a range of contingencies has led Hanson (1999) to term these 'decision markets'. Indeed, Hanson (2003a) has suggested that such markets could be used to remove technocratic policy implementation issues from the bureaucracy, a suggestion endorsed in Hahn and Tetlock (2006). Moreover, while the previous example involves only one contingency, Hanson (2003b) suggests that market scoring rules can allow traders to simultaneously predict many combinations of outcomes. The basic intuition of his proposal is that, rather than betting on each contingency, traders bet that the sum of their errors over all predictions will be lower.

However while contingent markets can be used to estimate the joint probability of choice A and outcome B, care must be taken before inferring that choice A should be made because it will maximize the probability of outcome B. That is, while these markets can highlight the correlation between events, the difficulty of inferring causation remains.



Prediction Markets, Fig. 5 Risk of war in Iraq. Prediction markets, export opinion and oil markets (*Source*: Trade-by-trade Saddam Security data provided by Tradesports.com; Saddameter from Will Saletan's daily column in Slate.com)

Conclusion

The healthy bibliography below attests to the fact that interest in prediction markets has boomed in recent years. Many questions remain. Theoretical research holds the promise of better understanding the institutional design features that yield optimal information aggregation and efficient pricing. The practical agenda includes developing new ideas about how and when prediction markets can aid decisionmaking by business and government.

See Also

- ▶ Capital Asset Pricing Model
- ▶ Cheap Talk
- ▶ Contingent Commodities
- ▶ Efficient Markets Hypothesis
- ▶ Forecasting
- ▶ Futures Markets, Hedging and Speculation
- ▶ Hedging
- ▶ Information Aggregation and Prices
- ▶ Noise Traders
- ▶ Terrorism, Economics of

Acknowledgment Thanks to Colin Camerer, Robin Hanson, Ryan Oprea, Charles Plott, Paul Rhode, and Erik Snowberg for very helpful comments.

Bibliography

Berg, J., R. Forsythe, F. Nelson, and T. Rietz. 2006. Results from a dozen years of election futures markets research. In *Handbook of experimental economic results*, ed. C. Plott and V. Smith. Amsterdam: North-Holland.

Berg, J., F. Nelson, and T. Rietz. 2003. *Accuracy and forecast standard error in prediction markets*. Mimeo: University of Iowa.

Berg, J., and T. Rietz. 2003. Prediction markets as decision support systems. *Information System Frontiers* 5: 79–93.

Berg, J., and T. Rietz. 2006. The Iowa Electronic Market: Lessons learned and answers yearned. In *Information markets: A new way of making decisions in the public and private sectors*, ed. R. Hahn and P. Tetlock. Washington, DC: AEI–Brookings Joint Center.

Camerer, C. 1998. Can asset markets be manipulated? A field experiment with racetrack betting. *Journal of Political Economy* 106: 457–482.

Chen, K.-Y. and Plott, C. 2002. Information aggregation mechanisms: concept, design and implementation for a sales forecasting problem. Working Paper No. 1131. Division of Humanities and Social Sciences, California Institute of Technology.

Grossman, S. 1976. On the efficiency of competitive stock markets where traders have diverse information. *Journal of Finance* 31: 573–585.

Grossman, S., and J. Stiglitz. 1976. Information and competitive price systems. *American Economic Review* 66: 246–253.

Gürkaynak, R., and J. Wolfers. 2005. Macroeconomic derivatives: an initial analysis of market-based macro forecasts, uncertainty, and risk. In *NBER international seminar on macroeconomics*, ed. C. Pissarides and J. Frankel. Cambridge, MA: NBER.

Hahn, R., and P. Tetlock. 2006. Using information markets to improve decision making. *Harvard Journal of Law and Public Policy* 29: 213–289.

Hanson, R. 1999. Decision markets. *IEEE Intelligent Systems* 14(3): 16–19.

Hanson, R. 2003a. *Shall we vote on values, but bet on beliefs?* Mimeo: George Mason University.

Hanson, R. 2003b. Combinatorial information market design. *Information Systems Frontiers* 5(1): 105–119.

Hanson, R. 2006. Designing real terrorism futures. *Public Choice* 128: 257–274.

Hanson, R., and R. Oprea. 2005. *Manipulators increase information market accuracy*. Mimeo: George Mason University.

Hanson, R., R. Oprea, and D. Porter. 2006. Information aggregation and manipulation in an experimental market. *Journal of Economic Behavior and Organization* 60: 449–459.

Herron, M., D. Cram, J. Lavin, and J. Silver. 1999. Measurement of political effects in the United States economy: a study of the 1992 presidential election. *Economics and Politics* 11: 51–81.

Knight, B. 2006. Are policy platforms capitalized into equity prices? Evidence from the Bush/Gore 2000 presidential election. *Journal of Public Economics* 90: 751–773.

Kyle, A. 1985. Continuous auctions and insider trading. *Econometrica* 53: 1315–1336.

Manski, C. 2004. *Interpreting the predictions of prediction markets. Working Paper No. 10359*. Cambridge, MA: NBER.

Milgrom, P., and N. Stokey. 1982. Information, trade and common knowledge. *Journal of Economic Theory* 26: 17–27.

Nelson, F., G. Neumann, and P. Polgreen. 2006. *Operating with doctors: Results from the 2004 and 2005 influenza markets*. Mimeo: University of Iowa.

Ortner, G. 1998. *Forecasting markets – an industrial application*. Mimeo: Technical University of Vienna.

Pennock, D., S. Lawrence, F. Nielsen, and C. Giles. 2001. Extracting collective probabilistic forecasts from web games. In *Proceedings of the seventh ACM SIGKDD international conference on knowledge discovery and data mining*. New York: ACM Press.
Plott, C. 2000. Markets as information gathering tools. *Southern Economic Journal* 67: 1–15.
Plott, C., and S. Sunder. 1982. Efficiency of experimental security markets with insider information: An application of rational expectations models. *Journal of Political Economy* 90: 663–698.
Plott, C., and S. Sunder. 1988. Rational expectations and the aggregation of diverse information in a laboratory security markets. *Econometrica* 56: 1085–1118.
Rhode, P., and K. Strumpf. 2004. Historical presidential betting markets. *Journal of Economic Perspectives* 18(2): 127–142.
Rhode, P., and K. Strumpf. 2006. *Manipulating political stock markets: a field experiment and a century of observational data*. Mimeo: University of North Carolina.
Roberts, B. 1990. Political institutions, policy expectations, and the 1980 election: A financial market perspective. *American Journal of Political Science* 34: 289–310.
Servan-Schreiber, E., J. Wolfers, D. Pennock, and B. Galebach. 2004. Prediction markets: Does money matter? *Electronic Markets* 14: 243–251.
Slemrod, J., and T. Greimel. 1999. Did Steve Forbes scare the municipal bond market? *Journal of Public Economics* 74: 81–96.
Snowberg, E., and J. Wolfers. 2005. *Explaining the favorite-longshot bias: Is it risklove, or misperceptions?* Mimeo: University of Pennsylvania.
Snowberg, E., J. Wolfers, and E. Zitzewitz. 2006. *Partisan impacts on the stockmarket: Evidence from prediction markets and close elections*. Mimeo: University of Pennsylvania.
Sunder, S. 1995. Experimental asset markets: A survey. In *Handbook of experimental economics*, ed. A. Roth and J. Kagel. Princeton: Princeton University Press.
Wolfers, J., and E. Zitzewitz. 2004a. Prediction markets. *Journal of Economic Perspectives* 18(2): 107–126.
Wolfers, J. and Zitzewitz, E. 2004b. Experimental political betting markets and the 2004 election. *The Economists Voice* 1(2). Online. Available at http://www.bepress.com/ev/vol1/iss2. Accessed 19 Feb 2006.
Wolfers, J., and E. Zitzewitz. 2005a. *Interpreting prediction market prices as probabilities*. Mimeo: University of Pennsylvania.
Wolfers, J., and E. Zitzewitz. 2005b. *Using markets to inform policy: The case of the Iraq war*. Mimeo: University of Pennsylvania.
Wolfers, J., and E. Zitzewitz. 2006. Five open questions about prediction markets. In *Information markets: A new way of making decisions in the public and private sectors*, ed. R. Hahn and P. Tetlock. Washington, DC: AEI–Brookings Joint Center.

Preference Reversals

Chris Starmer

Abstract

Preference reversal is a widely observed behavioural tendency for the preference ordering of a pair of alternatives to depend on the process used to elicit it. The phenomenon appears to be both a robust and a systematic departure from conventional preference theory. Competing theoretical explanations variously interpret it as a violation of procedure invariance (the presumption that preferences should be independent of the method of eliciting them); a failure of transitivity; or a consequence of loss-averse (and reference-dependent) preferences. This article discusses these interpretations, the related evidence, and reflects on some of the broader implications of the phenomenon.

Keywords

Allais paradox; Decision processes; Expected utility hypothesis; Expected utility theory; Intransitivity; Loss aversion; Preference reversal; Preferences; Procedure invariance; Regret; Savage's subjective expected hypothesis

JEL Classifications

C9

Preference reversal (PR) is a widely observed behavioural tendency for the preference ordering of a pair of alternatives to depend, in a predictable way, on the process used to elicit it.

The existence of preference reversal sets an empirical challenge to fundamental assumptions of conventional economic theory: PR is an apparent failure of procedure invariance (that is, the traditional presumption that preferences should be independent of the method of eliciting them).

Some see it as a challenge to the very idea that human decisions are governed by preferences.

Much of the empirical PR literature has examined decisions relating to pairs of simple gambles. One of the gambles (typically called the ‘P-bet’) will offer a relatively good chance of winning a modest prize, otherwise nothing (or sometimes a small loss); the other bet (the ‘$-bet’), offers a relatively small chance of winning a larger prize. In classic PR experiments, subjects are required to make straight choices between such pairs of bets and to provide separate (usually monetary) valuations for each bet. For any individual and gamble pair, conventional economic theory implies that the chosen gamble would also be the more highly valued of the pair. But while many individuals are so consistent, a significant proportion, typically, are not. The existence of some such inconsistency, by itself, is not especially surprising. People might, for instance, make a mistake in one or more task, leading to some level of inconsistency in comparisons of rankings. Interest in PR, however, stems largely from the fact that observed inconsistencies tend to be patterned in a highly predictable way: the typical finding is that considerable numbers of subjects choose the P-bet and value the $-bet more highly (let us call this the standard reversal), while very few commit the opposite reversal ($-bet chosen and P-bet valued more highly). It is this *asymmetric* pattern of inconsistencies between rankings based on choice and valuation that constitutes the intriguing PR phenomenon.

Evidence

PR was first predicted and then observed by psychologists (Lichtenstein and Slovic 1971; Lindman 1971). It was later brought to the attention of economists by Grether and Plott (1979) who described its potential significance for economics in the following passage:

> Taken at face value the data are simply inconsistent with preference theory and have broad implications for research priorities within economics. The inconsistency is deeper than mere lack of transitivity or even stochastic transitivity. It suggests that no optimisation principles of any sort lie behind even the simplest of human choices. (Grether and Plott 1979, p. 623)

Like many economists who have followed in their footsteps, Grether and Plott did not immediately accept this face-value interpretation and, instead, looked for ways of explaining PR while retaining the assumption that individuals do have a unique preference ordering over gambles. A substantial body of research in this spirit has examined whether PR might be an experimental artefact arising from imperfectly designed experiments. Early research of this genre – including Grether and Plott (1979), Reilly (1982) and Pommerehne et al. (1982) – investigated issues such as whether PR might be a consequence of subjects failing to understand the tasks confronting them, or of having insufficient motivation to take those tasks seriously. But a large body of evidence now shows that PR is a highly replicable phenomenon, robust to many variations in experimental procedures. Seidl (2002) provides a review.

A more subtle critique of PR experiments and evidence emerged in the late 1980s with the publication of a series of theoretical papers (Holt 1986; Karni and Safra 1987; Segal 1988) arguing that PR might be a spurious artefact of experimental design after all. These papers shared a common strategy, pointing to a potential weakness of two experimental procedures which had been commonly used to incentivize decision tasks in PR experiments: the (Becker et al. 1964) mechanism and the random lottery incentive system. The thrust of these papers is to show that, if individuals have non-expected utility preferences (violating either the independence axiom of expected utility theory, or the reduction of compound lotteries principle, or both), these standard incentive mechanisms could be biased and might generate the spurious appearance of PR. On this interpretation, PR would not be evidence against procedure invariance: instead it would be evidence of consistent, but non-expected utility, preferences interacting with specific features of experimental design. This interpretation has, however, been largely discounted in the light of subsequent research (including Tversky et al. 1990 and Cubitt et al. 2004) which reproduces the PR phenomenon in experiments using incentive mechanisms immune to this critique of earlier studies.

Theory

There remains considerable interest in trying to find a satisfactory explanation of PR. In what follows, we discuss three types of theory that may contribute to that objective: *regret theory, reference-dependent theory,* and *constructed preference theory.*

Regret theory (Loomes and Sugden 1982, 1983) explains PR as a form of intransitivity. In this theory preferences are defined over pairs of acts which map from states of the world to consequences (as in Savage 1954). Suppose A_i and A_j are two potential acts that result in, respectively, outcomes x_{is} and x_{js}, in state of the world s. If A_i is chosen, the resulting utility in each state is given by a 'modified utility function' $M(x_{is}, x_{js})$. Notice that this function allows the consequences of the chosen act to depend upon those that *might have been* experienced under the forgone act A_j. In particular, the utility from having x_{is} may be suppressed by 'regret' when x_{is} is worse than x_{js}. Regret theory assumes that individuals attempt to maximize the expectation of modified utility $\Sigma_s\ p_s.\ \mathrm{M}\ (x_{is}, x_{js})$ where p_s is the probability of state s. Regret theory reduces to expected utility theory in the special case where $(M\{x_{is}, x_{js}) = u(x_{is})$ and $u(.)$ is a von Neumann–Morgenstern utility function.

Loomes and Sugden (1982) show that, if preferences in this theory satisfy particular restrictions, then regret theory provides a possible explanation of several well-known violations of expected utility theory including some cases of the famous Allais paradox. The most important of these restrictions is a property (subsequently) called regret aversion and, in a follow-up paper, Loomes and Sugden (1983) show that regret aversion may also explain PR. The argument works roughly as follows. Consider the following three acts labelled $, P and M with monetary consequences x > y > m > 0 defined over three states.

	State1	State 2	State 3
$	x	0	0
P	y	y	0
M	m	m	m

The acts labelled $ and P have the structure of typical $- and P-bets: they are binary gambles where $ has the higher prize, and P the higher probability of 'winning'; the third act gives payoff m for sure. Regret theory allows choices over acts with this structure to be non-transitive and, if preferences are regret averse, if a cycle occurs it will be in a specific direction: P chosen over $; M over P; and $ over M. Now recall that, in a typical PR experiment, the standard reversal occurred when a subject chose P over $ but valued $ more highly than P. So, if we interpret choices from {$, M} and {P, M} as analogues of valuation tasks asking 'is $ (or P) worth more or less than M?', then the cycle predicted by regret theory can be interpreted as a form of PR.

This explanation for PR has been tested via experiments designed to look for the pure choice analogue of PR by confronting subjects with pairwise choices among triples of bets with the structure of $, P and M above. The outcome of this strand of research has produced good and bad news for regret theory. The good news is that the non-transitive choice cycles predicted by it have been observed and replicated (Loomes et al. 1991). Since these choice cycles occur in studies that involve no valuation tasks at all, this is evidence for the intransitivity interpretation of PR. The bad news is that subsequent research (Starmer and Sugden 1998) has cast considerable doubt on regret theory's account of these choice cycles. The current state of play appears to be that regret theory has led to the discovery of a surprising new choice phenomenon, but it turns out not to be the right explanation for it! It remains possible that these intransitive choice cycles are manifestations of regret-type influences at work but that formal models of regret must be refined to properly account for them. Another possibility is that they have nothing to do with 'regret' and that their discovery, as a consequence of testing regret theory, was just accidental.

A new account of PR has emerged in the form of reference-dependent subjective expected utility theory (Sugden 2003). In this model, preferences are again defined over acts. The key structural departure from Savage's (1954) subjective expected utility theory is that consequences in each state are modelled as gains and losses relative to a *reference act* (the status quo). The resulting theory is a formulation of expected utility (that is, a model that is linear in probabilities) that can accommodate loss aversion (that is, losses of a given size being weighted more highly than corresponding magnitude gains). Sugden demonstrates that, when preferences are loss averse, this model predicts standard PR in experiments where values are elicited as selling prices (which they usually are). This prediction depends on the assumption that, in selling tasks, an agent's reference act is the lottery being sold: given this, seemingly reasonable, assumption, $ valuations become particularly 'inflated' by consideration of the large $ prize which becomes a (probabilistic) loss if the $-bet is given up for a certain amount of cash. Hence, on this account, PR is the consequence of loss aversion operating through selling tasks. As yet, there have been no direct tests of this explanation, though the evidence of loss aversion operating in other contexts (see Starmer 2000, for some discussion) perhaps gives it some initial credibility.

Thus far we have discussed various preference-theoretic accounts of PR. The final type of explanation we discuss is the oldest and belongs to a class of theory that has evolved in the psychology literature. From the outset, most psychologists accepted PR as evidence against the very thing that economists have invested their efforts in defending: the presumption that behaviour can be adequately explained in terms of unique underlying preferences. Psychologists have, instead, focused on accounts of PR which attribute it to aspects of human *decision processes.* Viewed from this perspective, there is nothing fundamentally surprising about the fact that rankings delivered via choice and valuation tasks differ; those working within this paradigm will, typically, attempt to read such inconsistencies as clues to the, potentially distinct, mental heuristics invoked in those different tasks.

Numerous theories in this spirit have been proposed as putative accounts of PR, and one of the best known examples is the scale-compatibility hypothesis due to Tversky et al. (1988). The general hypothesis assumes that the way in which an individual is required to respond to a task ('the response mode') can affect the weights that he or she places on particular dimensions of alternatives being evaluated. In application to PR, the hypothesis implies that, because valuation tasks require a money amount as output, individuals place particularly high (low) weight on the money (probability) dimension, leading to relatively 'inflated' values for $ bets. Some recent support for this particular hypothesis is reported in Cubitt et al. (2004). There is, however, a vast theoretical and empirical literature connecting PR with the constructed preference approach and, for those interested in pursuing it, an excellent source is Lichtenstein and Slovic (2006).

Developing Themes

One developing theme in empirical PR research examines the persistence of PR in environments where individuals receive feedback on the consequences of their decisions. A famous experiment by Chu and Chu (1990) exposed preference reversers to 'money pumps': subjects who committed PR had their stated preferences implemented across a series of trades which ultimately resulted in monetary losses. Individuals quickly learned to avoid PR in this environment. While this is an interesting finding, since Chu and Chu use such an explicit method for disciplining inconsistent preferences, it would be a mistake to view this as persuasive evidence that PR would be eroded in any naturally occurring market. There is some limited evidence to suggest that PR may decay in some specific experimental markets (Cox and Grether 1996) but the findings here are both tentative and mixed, and further investigation is warranted before any firm conclusions can be drawn.

Another theme of current research explores the implications of preference anomalies (including PR) for the formulation of economic policy. A discussion of this topic is contained in Braga and Starmer (2005).

See Also

- ▶ Allais Paradox
- ▶ Expected Utility Hypothesis
- ▶ Learning and Evolution in Games: Adaptive Heuristics
- ▶ Paradoxes and Anomalies
- ▶ Prospect Theory
- ▶ Rational Behaviour
- ▶ Rationality
- ▶ Rationality, Bounded
- ▶ Savage's Subjective Expected Utility Model
- ▶ Transitivity

Bibliography

Becker, G.M., M.H. DeGroot, and J. Marschak. 1964. Measuring utility by a single-response sequential method. *Behavioral Science* 9: 226–232.

Braga, J., and C. Starmer. 2005. Preference anomalies, preference elicitation and the discovered preference hypothesis. *Environmental and Resource Economics* 32: 55–89.

Chu, Y.P., and R.L. Chu. 1990. The subsidence of preference reversals in simplified and marketlike experimental settings: A note. *The American Economic Review* 80: 902–911.

Cox, J.C., and D.M. Grether. 1996. The preference reversal phenomenon: Response mode, markets and incentives. *Economic Theory* 7: 381–405.

Cubitt, R.P., A. Munro, and C. Starmer. 2004. Testing explanations of preference reversal. *The Econometrics Journal* 114: 709–726.

Grether, D., and C.R. Plott. 1979. Economic theory of choice and the preference reversal phenomenon. *The American Economic Review* 69: 623–638.

Holt, C.A. 1986. Preference reversals and the independence axiom. *The American Economic Review* 76: 508–515.

Karni, E., and Z. Safra. 1987. "Preference reversal" and the observability of preferences by experimental methods. *Econometrica* 55: 675–685.

Lichtenstein, S., and P. Slovic. 1971. Reversals of preferences between bids and choices in gambling decisions. *Journal of Experimental Psychology* 89: 46–55.

Lichtenstein, S., and P. Slovic. 2006. *The construction of preference*. New York: Cambridge University Press.

Lindman, H.R. 1971. Inconsistent preferences among gambles. *Journal of Experimental Psychology* 89: 390–397.

Loomes, G.C., and R. Sugden. 1982. Regret theory: An alternative theory of rational choice under uncertainty. *The Econometrics Journal* 92: 805–824.

Loomes, G.C., and R. Sugden. 1983. A rationale for preference reversal. *The American Economic Review* 73: 428–432.

Loomes, G., C. Starmer, and R. Sugden. 1991. Observing violations of transitivity by experimental methods. *Econometrica* 59: 425–439.

Pommerehne, W.W., F. Schneider, and P. Zweifel. 1982. Economic theory of choice and the preference reversal phenomenon: A re-examination. *The American Economic Review* 73: 569–574.

Reilly, R.J. 1982. Preference reversal: Further evidence and some suggested modifications in experimental design. *The American Economic Review* 73: 576–584.

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

Segal, U. 1988. Does the preference reversal phenomenon necessarily contradict the independence axiom? *The American Economic Review* 78: 233–236.

Seidl, C. 2002. Preference reversal. *Journal of Economic Surveys* 6: 621–655.

Starmer, C.V. 2000. Developments in non-expected utility theory: The hunt for a descriptive theory of choice under risk. *Journal of Economic Literature* 38: 332–382.

Starmer, C., and R. Sugden. 1998. Testing alternative explanations of cyclical choices. *Economica* 65: 259–347.

Sugden, R. 2003. Reference-dependent subjective expected utility. *Journal of Economic Theory* 111: 172–191.

Tversky, A., S. Sattath, and P. Slovic. 1988. Contingent weighting in judgement and choice. *Psychological Review* 95: 371–384.

Tversky, A., P. Slovic, and D. Kahneman. 1990. The causes of preference reversal. *The American Economic Review* 80: 204–217.

Preferences

Georg Henrik von Wright

The concept of preference holds a pivotal position in value theory. It may even be considered a 'value radical' or common conceptual root of the three main types of evaluative discourse, namely, aesthetic, economic and moral.

In economics and the behavioural sciences generally, the role of this concept has recently been enhanced by the creation of 'exact' theories of 'strategic thinking' such as game theory, Bayesian decision theory, and a general theory of utility.

The concept of (individual) preference is not, on the whole, considered controversial by economists. Philosophical logicians, however, tend to regard the concept as problematic, and there is little agreement among them about the basic principles of a 'logic of preference'. The situation is a little like the one in probability theory. Probability is extensively used and successfully applied in both the physical and the social sciences – and yet philosophers notoriously disagree about its 'true meaning'.

The first systematic inquiry into the foundational problems of preferences seems to have been von Wright in 1963. It has at least one noteworthy precursor, Halldén (1957), which explores the related notion of 'betterness'. Significant contributions have been made later by R. Chisholm (1966, 1975), E. Sosa (1966) and N. Rescher (1967, 1969), and by the Swedes S. Danielsson (1968), P. Gärdenfors and B. Hansson (1968).

A statement of preference of the type which I shall call 'pure' or 'intrinsic' is a *value judgement.* It is *subjective* in the sense that it expresses somebody's preference of something over something else. It is *relative* in the sense that a subject's pure preferences may change in the course of time. Such changes can be spoken of as 'changes of taste'. (Another instance of relativity will be mentioned below.)

Sometimes we ask *why* a person prefers, say, *x* to *y.* And sometimes, not always, there is an answer at hand. The person can give a *reason* for his preference. For example: he prefers *x* to *y* as a means to the end *E* because *x* is, say, cheaper or quicker or safer than *y.* That *x* is cheaper (quicker, safer) is a factual statement, true or false as the case may be. That a person, other things being equal, prefers the cheaper means to the more expensive one is a valuation.

A preference for a reason I shall call 'extrinsic'. As seen from the example, an extrinsic preference is *linked* to an intrinsic one by means of an objective judgement (the reason). When the intrinsic preference is one which most people share, one often calls the preferred thing *preferable.* For example, the use of a safer means to a given end may be deemed preferable to the use of a less safe one. A judgement of preferability has an 'objective appearance'. It is an open problem in the philosophy of value whether *all* such judgements ultimately depend on subjective valuations expressed in pure preferences.

Sometimes the answer to the above Why? question is that the subject *likes x better* than *y.*

This is not to give a reason for the preference. It is a new verbalization of a pure preference. 'Liking better' is just another term for '(simply) preferring'.

The symbol '*xPy*' shall mean that *x* is preferred to *y,* subject and time left unspecified. For a 'logic' of the preference relation one could lay down the following axiomatic principles:

$$\begin{aligned}&\text{A1.}\quad xPy \rightarrow \sim(yPx)\\&\text{A2.}\quad xPy \rightarrow xPz \vee zPy.\end{aligned}$$

The first says that the *P*-relation is asymmetrical. The second says that if x is preferred to *y* then any third thing *z* is such that either *x* is preferred to it or it is preferred to *y.* This amounts, in effect, to saying that the *P*-relation is *connected,* that is, if two things are comparable for preference then any other thing may be compared to them.

From A1 and A2 one easily derives the following two theorems:

$$\begin{aligned}&\text{T1.}\quad xPy\,\&yPz \rightarrow xPz\\&\text{T2.}\quad \sim(xPx).\end{aligned}$$

They state that the *P*-relation is transitive and irreflexive.

We can define a relation of indifference as follows:

$$xIy =_{\mathrm{df}} \sim(xPy)\& \sim(yPx).$$

If the *P*-relation is assumed to obey A1 and A2, it may be proved that the *I*-relation is reflexive, symmetrical, and transitive. One can then say that

xIy means that x and y are of *equal* (intrinsic) *value* to the subject under consideration.

We can now also prove the theorems

$$\text{T3.}\quad xPy\,\&yIz \rightarrow xPz$$
$$\text{T4.}\quad xPy\,\&yIz \rightarrow zPy$$

They state that things of equal value are interchangeable in the preference relation.

The assumption that the *P*-relation is connected is a very strong assumption, the realism of which may be questioned. One might, for example, replace it by the weaker assumption of transitivity.

A *P*-relation which is (only) asymmetrical and transitive determines a *partial* ordering or ranking of alternatives. A *P*-relation which is in addition connected determines a *complete* ranking order. In a partial ordering, the *I*-relation is not provably transitive. Therefore it does not amount to value-equality. In a logic of partial orderings value-equality ('strong indifference') is a primitive concept. This concept cannot be defined in terms of preference (and negation), nor preference in terms of it.

The terms *x, y,*· · · of a *P*-relation can represent many different types of entity. They can be *goods,* for example when a person prefers apples to pears. Or they can be *means* to an end, for example travelling to a destination by bus rather than by train. Or *states of affairs* when, for example, revolt or war is preferred to continued oppression and slavery.

A person who professes a preference of some good *x* over another *y* presumably likes a state of affairs better in which he enjoys or possesses or uses or lives with *x* than one in which he has *y.* It may be held true that a relation of (pure) preference is basically a preference between two different states in which a person imagines himself to be.

Economists, it seems, usually treat the terms of the *P*- and *I*-relation as goods or other 'thing-like' entities, whereas logicians and philosophers tend to study them as states or otherwise 'proposition-like' entities. Technically, the second approach looks more interesting because it allows us to apply Boolean operations to the terms. As indicated, this may also be the more 'basic' approach.

For *P*-relations, the terms of which are 'thing-like', the assumption of connectedness seems, in general, too strong. A person prefers, say, (the taste of) apples to (the taste of) pears. He also prefers the music of Bach to that of Beethoven. But if asked whether he prefers apples to Bach or vice versa, his best answer is probably that he finds the alternatives 'incomparable'. By suitably limiting the *range* of things compared, one may, however, be able to secure that the preference order is complete and not only partial. Tastes of fruits, for example, may be throughout comparable for preference.

A theory of *P*-relations, the terms of which are states of affairs, needs some axiomatic principles in addition to those mentioned above. The new principles concern the 'behaviour' of the Boolean connectives in *P*-relations. On the details of the matter, however, there is widespread disagreement between researchers.

One could raise the question: What does it *mean* to prefer a state *x* to another one *y?* and answer: It means to think it better if the first state obtains but the second does not than the other way round, the second but not the first. One then accepts an equivalence $xPy \leftrightarrow x \;\&\; \sim y\; P \sim x \;\&\; y.$ From it one can easily derive a 'law of contraposition' for preferences, $xPy \leftrightarrow\; \sim y\; P \sim x$. Objections raised against it on intuitive grounds seem to me to confuse pure preferences between states with some other types of preference.

Another controversial question concerns the case when one or both terms of the relation are *disjunctive* states. An employee says he prefers an increase in salary *or* a shortening of his working day (salary remaining the same) to the status quo of his employment. Does this mean that he both thinks increased salary *and* also thinks reduced working hours preferable to his present position? To interpret the preference thus is to subscribe to an idea that a 'disjunctive preference' is resolvable into a *conjunction* of *P*-relations, the terms of which are not disjunctive.

The reader can easily satisfy himself that, taking 'contraposition' into account, a preference $x \vee yPz \vee u$ then is equivalent with a conjunction of

16 P-relations between 4-termed conjunctions of states and/or their negations.

A P-relation with the above properties is *holistic* in the following sense: A subject's preference of state x over state y obtains in a frame of 'accompanying circumstances'. Of great interest is the case when it obtains *ceteris paribus* or 'other things being equal'. Let these 'other things' be some conjunction of n states and/or their negations. There are in all 2^n such conjunctions or 'possible worlds', $C_1, \cdots, C_{2n}$. That x is preferred to y *ceteris paribus* means that every conjunction $x \,\&\, \sim y \,\&\, C_i$ is preferred to $\sim x \,\&\, y \,\&\, C_i$. xPy is then equivalent with the conjunction (totality) of the 2^n relations $x \,\&\, \sim y \,\&\, C_iP \sim x \,\&\, y \,\&\, C_i$.

The states x and y and the n states in C_i constitute the *preference horizon* of the person who has the preference. 'Ideally' this horizon should comprise every possible state in the world. In practice, however, it is limited to those states the possible relevance of which to his preference the subject happens to take into account.

Consider three P-relations xPy, yPz, and xPz. If we are to be able to infer the third from the first two, the preference must be taken relative to a preference-horizon which includes *at least* the states x, y, and z. xPy then means that x is preferred to y regardless of whether z or $\sim z$ obtains; yPz that y is preferred to z regardless of x; xPz that x is preferred to z regardless of y.

Thus xPy is explicated as $(x \,\&\, zPy \,\&\, z) \,\&\, (x \,\&\, \sim zPy \,\&\, \sim z)$. And similarly for yPz and xPz. It is easily seen that transitivity is then secured. If the three preferences had not fallen within one and the same 'horizon', transitivity need not have followed. It has sometimes been questioned whether the P-relation *is* (always) transitive. The answer is that transitivity is there only if due attention is paid to the preference-horizon.

The 'possible worlds' within a given preference-horizon may be ranked in a complete order of preference. But the ranking order of *all* the possible alternatives within such a horizon cannot be complete, but will have to be partial. This is a consequence of the fact that a conception of the P-relation as resolvable into a *conjunction* entails a conception of the I-relation as a *disjunction.* And this I-relation is not an equivalence (a value-equality). Thus a holistic conception of preferences between states entails that one must differentiate between 'mere' indifference and value-equality.

Must preference between states be conceived holistically? The opposite of a holistic conception is to consider the terms of the P-relation 'alone and taken by themselves'. This makes good sense when the terms are 'thing-like'. The taste of apples 'by itself' may be compared with the taste of pears 'by itself'. But whether a similar comparison makes good sense when the terms are 'proposition-like' (states) seems to me debatable. A non-holistic logic of intrinsic preferences between states has been proposed by Chisholm and Sosa (1966).

A question of interest to value theory is whether the *absolute* notions of goodness and badness can be defined in terms of the *relative* notion of betterness (preference).

Long ago, A.P. Brogan (1919) suggested that a state of affairs is good if it is better than its contradictory. In terms of preference:

$$\mathrm{Good}\, x =_{\mathrm{df}} xP \sim x.$$

Conversely, a state is bad if its contradictory is preferred to it.

If $xI \sim x$ holds good, x will be called indifferent *in itself.* If xIy holds, x and y will be called indifferent *between themselves.*

A weakness, among others, of the suggested definition of good is that one cannot prove that states which are indifferent in themselves are also indifferent between themselves. Thus it may happen that two states are indifferent in themselves, neither good nor bad, and one of them preferred to (thought better than) the other. This is counter-intuitive. Things which are neither good nor bad have no value and therefore should not be possible to rank as better or worse.

A similar objection can be made against the definition of goodness suggested by R. Chisholm (1975). It says that a state is good if it is preferred to some state which is indifferent in itself. But unless states indifferent in themselves are all equal in value it may happen that one and the same state is both good and bad. This also is counter-intuitive.

The upshot seems to be that the value-absolutes cannot be defined in terms of preference alone. They require in addition an independent concept of *zero*-value or (complete) valueneutrality.

One must distinguish between *preference* and (preferential) *choice.* A preference is an attitude, a choice an action.

If a subject has a pure preference for *x* over *y* it may be taken as analytic that, if he is offered a choice (option) between the two states, he will choose the former. It is then presupposed that the two states are being presented to him, so to say, 'on a tray' – that he can 'pick out' the one or the other without having to consider further prerequisites for or consequences of his choice.

From choices and options of this kind one must distinguish another type. I shall call these other options *conditional.*

A farmer is offered a choice between, on the one hand, getting a horse if it is raining tomorrow and a cow if it is not raining and, on the other hand, a cow if it is raining and a horse if it is not. He prefers getting a horse to getting a cow; this is a 'pure preference'. But which of the offered alternatives does he prefer? Assume that he professes to be indifferent as between them. How shall we then understand his attitude?

To this question there is an answer, first proposed by F.P. Ramsey, which has later come to play a great role in so-called Bayesian decision theory (Ramsey 1931; Savage 1954; Davidson 1955). Ramsey thought that an attitude of indifference here *means* that the person rates the two events, 'rain' and 'not rain', as *equally probable.* Accepting this, one can then proceed as follows:

Assume that our farmer is next presented with this option: On the one hand a horse if it is raining and a sheep if it is not raining and, on the other hand, a cow if it is raining and a hog if it is not raining. Again he says he is indifferent. This, on Ramsey's view, means that the value to him of a cow is *as much less* the value of the horse as the value of a sheep is less that of a hog. With this the way is open to a *metrization* of value and the introduction of *utility functions.* This done, one can use attitudes of indifference in other, more complex, conditional options for defining arbitrary degrees of (subjective) probability . The product of the value of a good and the probability of its materialization is called *expected utility.* Attitudes of preference in options aim at maximizing this quantity.

Ramsey's method is elegant and ingenious. Nevertheless, it seems to rest on a mistake. It ignores the distinction between the two senses of 'indifference'.

The farmer who, when presented with the first of the above two options, professes an attitude of indifference can do so for one of two reasons. Either he 'simply has no idea' about the chances of rainfall for tomorrow and therefore cannot make up his mind about which alternative is more to his advantage. This does not *mean* that he thinks rain and not-rain equally likely; he simply suspends judgement. *Or,* he considers them equally likely and *therefore* judges the two alternatives to be equally advantageous. He could, for example, support his attitude with the argument that if he repeatedly opted for one of the alternatives, no matter which one, on average half the number of times he would 'probably' get a horse, which is to his advantage, and half the number of times a cow, which is to his disadvantage. So therefore he is indifferent as between the alternatives. It is, in other words, not his judgement of indifference which gives meaning to the probabilities for him; but it is his prior estimate of the probabilities which determines his attitude of indifference. This estimate, moreover, seems normally to go with a corresponding expectation of frequencies.

P

The above criticism of Ramsey's procedure is not committed to a frequency theory of the 'meaning' of probability. But it assigns to expected frequencies a much more basic position for understanding probabilities than modern 'Bayesians' have tended to do (cf. von Wright 1962).

By a group-preference one can understand a ranking of alternatives in an order of ('objective') preferability based on the (subjective) rankings for preference of those alternatives by the members of the group. The derivation of the collective preference has to conform to some principles which seem intuitively plausible or 'rational'.

The problem of determining a group-preference is connected with notorious difficulties. Any proposed solution will depend partly upon which rational standards it is thought that the derivation should satisfy, and partly upon which demands are imposed upon the *P*- and *I*-relations involved.

In his influential book *Social Choice and Individual Values* (1951), K.J. Arrow showed that a derivation cannot collectively satisfy certain principles which individually seem plausible.

The result is known as Arrow's 'Impossibility Theorem'. It has been the topic of much subsequent discussion both by economists and logicians.

Group-preferences will not be further treated in this article. Let it be mentioned, however, that in the writer's opinion discussion has tended to neglect the complications connected with the concept of *indifference*. It is usually assumed that the *P*-relation is, at least, transitive and that the *I*-relation is an equivalence relation. In a preference ranking which is a partial ordering, this requirement on the *I*-relation is not automatically fulfilled. And even if the *I*-relation for individual preferences were an equivalence, it may not be 'reasonable' to demand or expect the *I*-relation for a group-preference to be this. Observing the distinction between 'mere' indifference and value-equality may therefore be helpful in efforts to cope with the conceptual difficulties in this area.

See Also

- ▶ Characteristics
- ▶ Ramsey, Frank Plumpton (1903–1930)
- ▶ Representation of Preferences
- ▶ Social Choice
- ▶ Utility Theory and Decision Theory

Bibliography

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

Brogan, A.P. 1919. The fundamental value universal. *Journal of Philosophy* 6(4): 967–1104.

Chisholm, R.M. 1975. The intrinsic value of disjunctive states of affairs. *Noûs* 9(3): 295–308.

Chisholm, R.M., and E. Sosa. 1966. On the logic of 'intrinsically better'. *American Philosophical Quarterly* 3(3): 244–249.

Danielsson, S. 1968. *Preference and obligation*. Uppsala: Scriv Service.

Davidson, D., J.C.C. McKinsey, and P. Suppes. 1955. Outlines of a formal theory of value. *Philosophy of Science* 22(2): 140–160.

Halldén, S. 1957. *On the logic of 'better'*. Lund: Gleerup.

Hansson, B. 1968a. Fundamental axioms for preference relations. *Synthese* 18(4): 423–442.

Hansson, B. 1968b. Choice structures and preference relations. *Synthese* 18(4): 443–458.

Hansson, B. 1969a. Group preferences. *Econometrica* 37(1): 50–54.

Hansson, B. 1969b. Voting and group decision functions. *Synthese* 20(4): 526–537.

Hansson, B. 1970. *Preference logic, philosophical foundations and applications in the philosophy of science*. Lund: Studentlitteratur.

Houthakker, H.S. 1965. On the logic of preference and choice. In *Contributions to logic and methodology in honor of J.M. Bocheński*, ed. A.-T. Tymieniecka. Amsterdam: North-Holland.

Ramsey, F.P. 1931. Truth and probability. In *Foundations of mathematics and other logical essays*, ed. F.P. Ramsey. London: Kegan Paul.

Rescher, N. 1967. Semantic foundations for the logic of preference. In *The logic of decision and action*, ed. N. Rescher. Pittsburgh: Pittsburgh University Press.

Rescher, N. 1969. *Introduction to value theory*. Englewood Cliffs: Prentice-Hall.

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

von Wright, G.H. 1962. The epistemology of subjective probability. In *Proceedings of the 1960 International Congress: Logic, methodology and philosophy of science*, ed. E. Nagel, P. Suppes, and A. Tarski. Stanford: Stanford University Press.

von Wright, G.H. 1963. *The logic of preference*. Edinburgh: Edinburgh University Press.

von Wright, G.H. 1972. The logic of preference reconsidered. *Theory and Decision* 3(2): 140–169.

Preindustrial Inequality

Branko Milanovic

Abstract

This article considers inequality in pre-industrial societies, defined as those prior to the industrial revolution and subsequent non-industrial societies that are not systematically integrated into the advanced world's economy. Although data on individual incomes and wealth in these societies are limited, increasingly they are becoming available. On the basis of these data, inequality as measured by the Gini coefficient is often on a par with modern industrialized societies, but the income gradient tends to be different, with a mass of people at subsistence level or marginally above, few at the mean, and a small affluent class. More work remains to be done, particularly on the relationship between income inequality and economic progress.

Keywords

Gini coefficient; Income distribution; Inequality; Kuznets curve; Preindustrial societies

JEL Classifications

D31; N30; O1

Defining Preindustrial

We need to circumscribe the scope of preindustrial. At one level, it is easy: preindustrial economies are characterized by low urbanization rates, high share of agriculture in GDP, low literacy rates, and of course low overall GDP per capita. However, many of today's poor countries share precisely these features. They are however 'non-industrial' or 'non-industrialized' rather than 'preindustrial' economies: this is because they are part of the modern world, systematically included in trade and voluntary movements of factors of production ('globalization') and have social structures which are very different from those of preindustrial societies. The life expectancy of their populations as well as the immunization and school enrolment rates exceed many times those of 'true' preindustrial societies. Not the least important is the fact that political compulsion of slave or serf labour, so ubiquitous in all preindustrial societies, is – except in a few pockets – largely absent.

Our definition of preindustrial includes all societies prior to the industrial revolution, and those that have not engaged with the industrial revolution, only up to a point when they began to be integrated systematically, rather than episodically, into the world economy. For many of them, integration coincides with colonization. Thus, broadly speaking – since we are painting with a very broad brush here – we can set limits around the end of the Napoleonic wars for Western Europe and the United States and Canada, and the end of the 19th century for everybody else. Twentieth-century societies, even when poor and hardly industrialized, belong to a different category.

A cut-off date around 1815–20 is convenient for at least three reasons. Politically, it coincides with a 'rearrangement' of Europe and, as later emerged, the world. It marks the beginning of the 'long 19th century'. Economically, it marks, according to the new English wage data series produced by Clark (2005), the beginning of a long-run rise in real wages which is continuing to this day. In terms of history of economic thought, Ricardo's *Principles* were published in 1817.

An obvious, but nevertheless important, clarification is that we are concerned here with income inequality: that is, inequality that includes all sources of income and reflects differences in households' and individuals' living standards. This, for example, rules out wage or rural–urban inequalities as such. (Wage inequality has meaning only if calculated across all wage-earners; income inequality includes the entire population.)

Implicit Theory

We do have an implicit theory about income inequality in preindustrial economies. The Kuznets hypothesis (formulated in 1955), the bread and butter of inequality economics, posits that inequality charts an inverted U shape as economy transforms from predominantly agricultural to predominantly industrialized or modern. In Kuznets' own words:

> One might thus assume a long swing in the inequality characterizing the secular income structure: widening in the early phases of economic growth when the transition from the preindustrial civilization was

> most rapid, becoming stabilized for a while; and then narrowing in the later phases. (Kuznets 1955, p. 276)

The same hypothesis, albeit without the mechanism that generates the inverted Ushaped curve, was formulated 120 years before Kuznets by Tocqueville:

> If one looks closely at what has happened to the world since the beginning of society, it is easy to see that equality is prevalent only at the historical poles of civilization. Savages are equal because they are equally weak and ignorant. Very civilized men can all become equal because they all have at their disposal similar means of attaining comfort and happiness. Between these two extremes is found inequality of condition, wealth, knowledge-the power of the few, the poverty, ignorance, and weakness of all the rest. (de Tocqueville 1835, pp. 42–3)

From both we should retain the sense that inequality is supposed to emerge only when societies are richer, and thus inequality in preindustrial societies may be expected to be low. But differently, we also have an image of preindustrial societies as combining abject poverty in the bottom with extravagant wealth on the top. For example, in ancient Rome, Goldsmith (1984, p. 287) notes the extraordinarily high income of the rulers relative to Great Britain in the early 19th century. Could both these images be right? As we shall argue below, yes – and this is one of the key features that distinguishes inequality in premodern times from inequality in modern times.

But in order to speak about inequality in preindustrial societies, we must also assume that preindustrial societies were 'modern' in the sense that they were (predominantly) market-oriented economies with non-negligible monetized sectors – and when they were non-monetized, goods and services given or received for political or power reasons could be valued at some meaningful 'market' prices. This is a position not universally accepted. In a famous debate about the later Roman Empire (and, by extension about all ancient economies) and 'modernity', there were two camps: that of 'primitivists' led by Polanyi (1944), Finley (1985) and Schiavone (1995), and that of 'modernists' (Rostovtzeff 1926; Walbank 1946). The first believed that Rome lacked most of the modern concepts that we associate with a market economy. Market relations, even when present, were of peripheral importance, and a market economy, itself a recent phenomenon, is perhaps, in a historical sense, only a brief episode (Polanyi 1944). For the 'modernists', the links between a preindustrial society like Rome and modern capitalism were obvious. Both Rostovtzeff and Walbank write of Roman 'bourgeoisie'. Whatever our opinion about the respective merits of 'primitivists' and 'modernists', it is important to realize that once we attempt to make some tentative estimates of economic inequality in preindustrial societies, we *ipso facto* accept that, while preindustrial societies might have been poorer and with a different social structure from modern societies, the differences are of magnitude, not of kind. For if such key concepts of market economy as prices, wage-labour and private property are vague, insufficiently understood by the population, not sanctioned by custom or law, then applying modern economic categories may be meaningless. Every attempt to study preindustrial societies empirically using today's economists' tools must assume that 'ancient' and 'modern' are fundamentally the same – so that that the 'ancient' can be described and understood using economic concepts developed from Adam Smith onwards.

Private property must enter the list above with a caveat. No one would deny that socialist societies, where private property was limited, were not modern. Moreover, they regarded themselves as the epitome of modernity. Similarly, societies with largely communal ownership of land (as in Africa) are modern too. Thus, private property of the means of production seems to be less of a requirement for a modern society than for example monetization. Rawls (1971), who can hardly be seen as a non-modernist, allows in his *Theory of Justice* for both private and nonprivate ownership of the means of production (see pp. 54, 240–1).

Data for Preindustrial Inequality

Where do data for preindustrial inequality come from? Since the Second World War, empirical studies of income distribution have been based on household surveys (nationally representative samples of households who are anonymously interviewed about their household characteristics, spending patterns and income). The earliest household surveys are from late 18th-century England. There were a few sporadic surveys in the 19th century (continental Europe, rural Russia) but they spread broadly only after the end of the Second World War, and as far as Africa and China are concerned, surveys became available only more recently, from the early 1980s. Obviously, such surveys were not conducted in any preindustrial society – even if censuses (driven by government tax needs) were. However, there are relatively abundant sources that economists can use to gauge income distribution in preindustrial societies, although the sources are often buried in hard-to-access archives and books, written in languages and alphabets that are not widely known, and requiring large amounts of both money and effort to be brought to light in a usable form. (For example, Ottoman censuses are written in Turkish but using Arabic script, rather than Latin as is used in today's Turkish. To process them requires knowledge of an often archaic Turkish and an alphabet into which this language is no longer written. See Cosgel 2002, 2004.) And then lots of heroic assumptions are needed in order for them to be 'translated' into modern economic categories. This has severely limited the use of ancient sources, and this is probably why only a fraction of such sources has been used so far.

The most comprehensive contemporary sources are tax data and government censuses undertaken in order to supply governments with information about taxation and the war-waging capacity of the populace (number of men, houses, horses, grain). Early documentary evidence includes government edicts (such as Diocletian's edict on maximum prices and wages from 301, recently studied by Allen 2007), as well as numerous Roman papyri preserved in the dry climate of Egypt. The English Domesday survey of 1086 is perhaps the best known of such sources.

From the Byzantine Empire, we have a few preserved *praktika* that provide descriptions of household characteristics, inventories of possessions and taxes paid, although they cover only limited areas (towns or ecclesiastical communities). (See the multi-volume *Economic History of Byzantium: From the Seventh through the Fifteenth Century* edited by Angeliki Laiou 2002.) Ottoman censuses (*defterlar*) from approximately the 14th century onward, conducted to assess the wealth and military capacity of newly conquered territories, provide detailed information on settlements (hamlets, villages, small and larger towns) but then present it in average amounts for each settlement (not by individual household). If inequality within settlements is not huge, and the number of settlements included is large, censuses can be used to assess overall income distribution within a country or a region.

A much-used source is the Florentine *Catasto* from 1427. (The data were originally collated by Herlihy and Klapisch-Zuber 1985. Currently, they are available on the Internet.) The Spanish *Ensenada Cadastre*, similar to modern-day household surveys, was carried out in the 1750s for the purposes of a neverimplemented fiscal reform. It has recently been used by researchers, and will be no doubt analysed more once it is digitized. Inequalities for the cities of Paris, Amsterdam and London were studied from tax data for respectively 1292–1313 (Sussman 2005), 1732–42 (McCants 2007; Soltow 1989) and 1797–1801 (Schwarz 1979). However, they refer to wealth inequality (there is no attempted 'conversion' to income), cover very truncated data sets, focus either on the rich – those subject to taxation – or the poor (McCants 2007), and of course include single cities only. Incidentally, all examples but one used by Pareto in the formulation of his famous 'iron law' of income distribution come from various European tax data from the end of the 19th century (see Pareto 1896). The data on Latin America, produced by various Spanish *Visitas*, which collected detailed information on population, age, land ownership and agricultural output, have been published in numerous volumes but not used for estimates of income distribution. (For Peru, books with detailed notes from *Visitas* for the years 1562, 1567 and 1604–05 have been published.)

What is common to these sources is that they are in principle surveys of stocks (people and wealth) and require a huge effort of price imputation; first, to 'transform' a stock into a meaningful annual yield (income), then to convert produced quantities, expressed in local 'natural' units (such as Egyptian *modii* of wheat), into kilograms, and finally to convert all of these into monetary units. Then the researcher needs to resort to even more heroic assumptions to calculate other sources of income, from husbandry, vineyards, honeybee cultivation, fruits and plants, services provided by farmers, and not least, from manufacturing activities like pottery, glass or clothmaking, or provision of urban services from the shoe-maker to the teacher (for which, at least some wage data are generally available). Particularly vexing is the issue of measurement units, volumes or weights with often confusingly similar or identical names, which nevertheless imply different physical amounts from one region to another; or when money units are provided, the issue of silver or gold conversion between them. But such sources, however frustrating. can and do provide very useful evidence about ancient living standards and distribution of income.

The second contemporary evidence is provided by social tables. This is what William Petty termed 'political arithmetick'. They aim to describe the structure of a society by listing all salient social classes (or professions) and estimating their average incomes (per household, or less often per person). For modern economists, these sources are much easier to use because the classification into presumably socially important groupings and estimates of their money-equivalent incomes provide us with most of what we need to know for the derivation of income distribution. England was the pioneer in the production of social tables, beginning with the famous one of Gregory King for 1688 (which contains 33 social groups with their population sizes and average incomes), and continuing with Massie (1759) and Colquhoun (1801–3). (None of the social tables, or the results obtained from them, is without its critics: for a critique of King's social table, see Arkell 2006; for Colquhoun, see Schwarz 1979; for a critique of Lindert–Williamson's use of English social tables, see Feinstein 1988.) Much more recent authors have produced similar social tables for a number of countries (see, for example, Morrisson and Snyder 2000, for France in 1788, Bértola et al. 2006, for Brazil in 1872, van Zanden 2003, for Java in 1880, Berry 1990, for Peru in 1876). These new social tables are of course not contemporary sources but they were produced, using bits of dispersed primary or most often secondary sources, by economic historians who specialize in various eras and countries, and they represent our best guess at social structure at remote points in time. The work of Milanovic et al. (2009; hereafter MLW), who made the first systematic attempt to measure and analyse preindustrial inequality, is largely based on such (contemporary and recent) social tables.

Empirical Evidence

To translate preindustrial inequality into modern economics, we must not only hold that preindustrial societies were largely monetized (and whatever was not monetized could be ultimately expressed in money), but also hold that their inequality can be meaningfully handled by Gini, Theil or any other currently used inequality measure. Otherwise we lack a common yardstick with which to compare past and present.

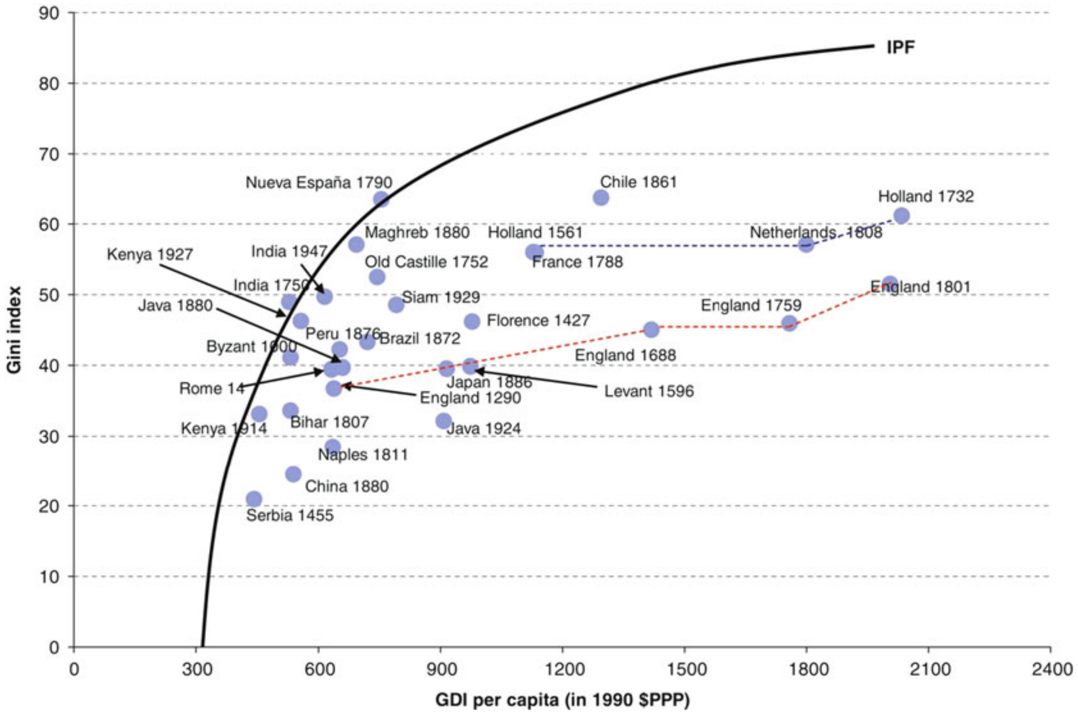
Using mostly social tables from 30 preindustrial societies, MLW calculated Gini coefficients. They found that the preindustrial Ginis range from the mid-20 s to around 65, with a mean of 45 and standard deviation of 11. (Gini is the most commonly used measure of inequality, and ranges from a theoretical zero (everybody has the same income) to a theoretical maximum of 100 (everybody but one person has a zero income, and the richest person takes the entire income of the community).) This is almost the same as the range of Ginis in modern societies. In fact, the modern equivalents of the preindustrial societies included in MLW sample (such as Turkey for Byzantium, Syria for the Levant, today's United Kingdom for the 1688 England and Wales) have an average inequality of 40 Gini points with a standard deviation of 10. However to make such a simple comparison and leave it at that would be erroneous. Preindustrial and modern societies were very different, even when compared in the language of modern economics.

First, it is very likely that the income gradient (how income increases as we move from poorer to richer income classes) was much flatter in preindustrial that in modern economies (see MLW 2009). Using Jan Pen's (1971) metaphor of dwarfs and giants, where people are visualized as marching in a 60-minute parade, from the poorest to the richest, with everyone's height reflecting their income, preindustrial societies can be seen as societies of dwarfs who would take some 40 to 45 minutes to file past. They contained large groups of people (most of the time, the vast majority of the population) living at, or just above, the subsistence minimum. Percentage differences in income among this vast mass of people were small. The income gradient was flat up to a very high point in income distribution. But then, and quickly, as we approach the very end of the parade, the gradient would suddenly increase, much more so than in modern societies. Thus, unlike a modern parade which would be characterized by a steady increase of the gradient, in preindustrial societies the middle was not much different from the bottom. There was a dearth of people whom we would (using modern terminology) identify with the middle class. (It is worth pointing out that this 'middle class' is not defined in terms of absolute income, or what we would consider today to be middle-class requirements, but entirely in terms of the period average income.) We can thus see why both of our preconceived notions – of generalized equality and drastic income disparity among the ancient – are true: they just refer to different parts of income distribution.

This difference in structure implies that the same calculated measures of inequality have different meanings. Ginis, as we have already indicated, were broadly in the same range then and now. But a Gini of 40, estimated independently for the Roman Empire by MLW (2009) and Scheidel and Friesen (2009), had an altogether different meaning from the same Gini in the contemporary United States. (The MLW estimate refers to the year 14 (at the death of Octavian), Scheidel's estimate to the mid-second century.) The Roman Empire's mean income was about twice the physiological subsistence level (*s*). If we require that all members of a society have at least the subsistence minimum – for otherwise the society will tend to shrink and disappear – then a very low level of mean income, regardless of how tiny the upper class is, limits the extent of measured inequality. Simply put, the extent of inequality is limited by the size of average income. That ceiling is more binding when a society is poor. To realize this, assume that society's mean income is just a fraction above *s*. If all but a tiny elite live merely on *s*, the elite cannot be extravagantly rich because total income is low, and Gini or Theil indexes, which take into account incomes differences between all individuals, cannot be very high either. This is the idea underlying the Inequality Possibility Frontier (IPF: see Fig. 1), defined by MLW (2009) and Milanovic (2006).

The frontier gives a maximum Gini (or Theil) coefficient which is compatible with a given level of mean income and maintenance of society as a going concern. The maximum Gini is equal to $(a-1)/a$ where $a =$ mean income divided by *s*, or the number of subsistence minima contained in the mean. As can be seen from the formula, the maximum feasible Gini rises with mean income (a), but at a decreasing rate. If average income is twice the subsistence ($a = 2$), the maximum Gini will be 50. Thus, we see that the Roman inequality of 40 exhausted some 80 per cent of maximum feasible inequality. But for the modern-day United States, where the mean income stands at more than 100 *s*, the maximum Gini is 99. The actual inequality will have exhausted only 40 per cent of its maximum value. Hence, the social meaning of the same Gini is entirely different. To sustain high inequality, societies must be relatively rich.

We have left the issue of defining the subsistence minimum deliberately vague. Depending on whether we pitch this physiological (note: not social, not relative) minimum higher or lower, the IPF will move down or up, but the same logic will hold.



Preindustrial Inequality, Fig. 1 The Inequality Possibility Frontier (*Source*: MLW (2009))

The difference in the income structure (income gradient) also shows why some other measures, like top-to-bottom ratio or top 1 per cent share, may not be very useful in the preindustrial context. They show the extent of the gap between the richest and the poorest, but they disregard the entire distribution in-between, which in the past has been much more equal than in today's societies.

IPF imposes a consistency check on our inequality calculations, a fact which is particularly useful for preindustrial societies where the evidence is scant. As illustrated in the figure, once we know the mean income of a society, and estimate its Gini, we know that this estimate must be within, or at the maximum on, the frontier. If it is not, there is something wrong with either the income or the inequality estimate, or the society is doomed to experience a dwindling population and ultimately extinction. It is not surprising that MLW found that all six cases of ancient societies with inequalities close to the frontier were colonies: India in 1750 and 1947, Kenya in 1914 and 1927, Nueva España (Mexico) in 1790, and Maghreb in 1880. Colonizers were clearly much less concerned about the welfare of the populations they ruled than, or did not have to fear them as much as, native rulers.

Preindustrial Inequality and Modern Debates

Empirical evidence on preindustrial inequality has a direct bearing on several contemporary debates. Evidence from the two most advanced economies at the time (England and Holland) paints a picture of increasing inequality from 16th century to the beginning of the Napoleonic wars. (The exception is Soltow 1968, who found English inequality to have been flat throughout the 18th century.) Premodern growth seem to have exacerbated inequality even in the areas that were characterized by an already high inequality of wealth and income (such as the South Midlands in England, considered by Allen 1992). Using social tables, Lindert (2000) and Lindert and Williamson (1982, 1983) document the increase of inequality in England between 1750 and 1801. All four observations available for England and Wales in the MLW database (1290 – from Campbell 2007–1688, 1759 and 1801–3) show both mean income and inequality rising with time. Similarly, van Zanden (1995), and Soltow and van Zanden (1998) find that income inequality increased in Holland during its 'Golden age': between 1561 and 1732: the urban area Gini rose from 53 to 59, and the rural area Gini from 35 to 38. According to a pioneering study by Hoffman and colleagues (2002), 'real' European inequality between 1500 and the early 19th century increased even more because the prices of wage-goods, consumed by the poor, rose relative to the prices of 'luxuries'.

The upswing of the Kuznets curve seems to be strongly in evidence in all these cases. But what drove it? Was it a 'classical explanation' (as van Zanden 1995, terms it), namely a shift in the functional distribution of income toward property owners (and their rising concentration) and away from labour – a mechanism that Marx would easily have recognized? (For Spain, Prados de la Escosura 2008, uses functional distribution of income, and also finds a clear Kuznets upswing from 1850 to around 1914.) Or was it, as argued by Lindert and Williamson (1985) and Williamson (1982, 1985), caused by the 'wage-stretching' which continued well into the 19th century and involved labour-saving technological progress and increased pay-ratios for skilled labour in the presence of demographic pressure from mostly unskilled population? Education responded only very slowly, and the process continued for a couple of centuries until massive European emigration reversed it. The latter is a very neoclassical mechanism familiar to every economist working on poor or rich countries today. The focus is on the functioning of factor markets, not on the division of society into capitalists and workers.

If countries where the industrial revolution originated went through a period of sustained increase in inequality prior to the industrial revolution, does it shed some light on the relationship between higher inequality and the industrial take-off? A number of recent writings (most famously, Pomerantz 2000; Frank 1998; and more recently Wen 2009; Shiue and Keller 2007) have contrasted China and Western Europe in the 17th and 18th centuries, trying to understand why these two large areas that seemed in many respects similar (for instance in market integration, level of income, technological innovations) charted such different paths in the following three centuries. Does income distribution have to do something with it? Unfortunately, we do not yet have even the intimation of an answer because the historical data for China are not available. However a recent upsurge in archival research on Chinese sources might help throw some new light on this issue.

The work of Engerman and Sokoloff (1997) has profoundly affected our conception of the role of inequality in explaining the economic success of North America and relative decline of Latin America. But while there is little doubt that Latin America was more unequal (particularly in land ownership) that the North, recent historical evidence contrasting Western Europe and Latin America finds no perceptible difference in inequality between the two. Williamson (2009) thus wonders why Western Europe and Latin America have followed different growth trajectories. If the inequality explanation works for one set of regions (the two 'New Worlds'), why does it seem not to work for another (Europe and Latin America)? Moreover, it is not evident that Latin America was 'always' unequal. Prados de la Escosura (2007) and Bértola et al. (2009) argue that strong expansion of inequality occurred during the previous round of globalization (1870–1920). Prados de la Escosura (2007, p. 298) sees the explanation as consistent with the factor-price equalization theorem: opening up Latin America to trade raised land rents, and since land was unequally distributed, increased the concentration of incomes. The data prior to around 1870 are not available (although some estimates for 1870 show inequality in the Southern Cone countries to be at the same level as in Spain: Prados de la Escosura 2008, Fig. 8, p. 307), but we could wonder whether our 'acquired idea' of an always high inequality in Latin America is not mistaken – or perhaps it was not inequality,

P

but the inequality extraction ratio that was high. Recasting the issue in this way suggests that the Latin American problem was a low level of income rather than a high Gini.

Conclusion

Studying inequality in its historical context, an area which will doubtlessly loom larger in economics as the search to uncover our economic past progresses, is important not only because it helps us learn about history but because it helps us understand today's economic problems. Actually, as every historian and politician knows, studying the past is about the future.

See Also

- ▶ Economic History
- ▶ Inequality (International Evidence)
- ▶ Inequality (Measurement)
- ▶ Standards of Living (Historical Trends)
- ▶ Wage Inequality, Changes In

Acknowledgments I am grateful to Mihail Arandarenko, Peter Lindert, Leandro Prados de la Escosura and Jeffrey Williamson for excellent comments.

Bibliography

Allen, R.C. 1992. *Enclosures and the Yeoman: Agricultural development of the South Midlands, 1450–1850*. Oxford: Oxford University Press.

Allen, R.C. 2003. Progress and poverty in early modern Europe. *Economic History Review* 56: 403–443.

Allen, R.C. 2007. *How prosperous were the Romans? Evidence from Diocletian's Price Edict (AD 301)*, Discussion Paper Series 363. Oxford: University of Oxford, Department of Economics, October.

Arkell, T. 2006. Illuminations and distortions: Gregory King's scheme calculated for the year 1688 and the social structure of later Stuart England. *Economic History Review* 59: 32–60.

Berry, A. 1990. International trade, government and income distribution in Peru since 1870. *Latin American Research Review* 25(2): 31–59.

Bértola, L., C. Castelnovo, E. Reis, and H. Willebald 2006. Income distribution in Brazil, 1839–1939. Paper presented at Session 116 of "A Global History of Income Distribution in the Long 20th Century," *XIV International Economic History Congress*, Helsinki-Finland 21–25 August.

Bértola, L., C. Castelnovo, J. Rodriguez, and H. Willeband 2009. Income distribution in Latin American Southern Cone countries during the first globalization boom, ca. 1870–1920. *International Journal of Comparative Sociology*, forthcoming.

Campbell, B. 2007. Benchmarking medieval economic development: England, Wales, Scotland and Ireland c. 1209. *Economic History Review* 60: 1–50.

Clark, G. 2005. The condition of the working class in England, 1209–2004. *Journal of Political Economy* 115: 1307–1340.

Cosgel, M. 2004. Ottoman tax registers. *Historical Methods* 37(2), Spring, 87–100.

Cosgel, M. 2006. Taxes, efficiency, and redistribution: Discriminatory taxation of villages in Ottoman Palestine, Southern Syria, and Transjordan in the sixteenth century. *Explorations in Economic History* 43: 332–356.

de Tocqueville, A. 1835. *Memoir on pauperism*, 1997. Chicago: Ivan R. Dee.

Engerman, S., and K. Sokoloff 1997. Factor endowments, institutions and differential paths of growth among New World economies. In *How Latin America fell behind: Essays on the economic histories of Brazil and Mexico, 1800–1914*, ed. S. Hager, 260–304. Stanford, CA: Stanford University Press.

Feinstein, C. 1988. The rise and fall of the Williamson curve. *Journal of Economic History* 48: 699–729.

Finley, M. 1985. *The ancient economy*, 2nd ed. London: Penguin.

Frank, A.G. 1998. *Re-orient: Global economy in the Asian age*. Berkeley: University of California Press.

Goldsmith, R.W. 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.

Herlihy, D., and C. Klapisch-Zuber. 1985. *Tuscans and their families*. New Haven: Yale University Press.

Herlihy, D., C. Klapisch-Zuber, R.B. Litchfield, and A. Molho *The online catasto*. Available at: http://www.stg.brown.edu/projects/catasto/overview.html. Accessed January 2008 (a searchable online database of tax information for the city of Florence in 1427–29, based on D. Herlihy and C. Klapisch-Zuber, principal investigators, Census and property survey of Florentine dominions in the province of Tuscany, 1427–1480).

Hoffman, P.T., D. Jacks, P.A. Levin, and P.H. Lindert. 2002. Real inequality in Europe since 1500. *Journal of Economic History* 62: 322–355.

Kuznets, S. 1955. Economic growth and income inequality, presidential address to the 67th meeting of the American Economic Association, Michigan, December 1954; published in *American Economic Review* 45 (1955) and *Economic growth and structure: Selected essays*. New Delhi: Oxford and IBH, 1965.

Laiou, A. 2002. *The economic history of Byzantium: From the seventh through the fifteenth century*. Washington, DC: Dumbarton Oaks.

Lindert, P.H. 2000. Three centuries of inequality in Britain and the United Stares. In *Handbook of income distribution*, ed. A. Atkinson and F. Bourguignon. Amsterdam: Elsevier.

Lindert, P.H., and J.G. Williamson. 1982. Revising England's social tables, 1688–1812. *Explorations in Economic History* 19: 385–408.

Lindert, P.H., and J.G. Williamson. 1983. Reinterpreting Britain's social tables, 1688–1913. *Explorations in Economic History* 20: 94–109.

Lindert, P.H., and J.G. Williamson. 1985. Growth, equality and history. *Explorations in Economic History* 22: 341–377.

Malanima, P. 2006. Pre-modern equality: Income distribution in the Kingdom of Naples (1811). Paper presented at *14th International Congress of Economic History*, August 2006, Helsinki. Available at: http://www.helsinki.fi/iehc2006/papers3/Malanima.pdf.

Mayhew, N.J. 1995. Modeling medieval monetization. In *A commercialising economy: England 1086 to c. 1300*, ed. R.N. Britnell and B.M.S. Campbell, 55–77. Manchester: Manchester University Press.

McCants, A. 2007. Inequality among the poor of eighteenth century Amsterdam. *Explorations in Economic History* 44: 1–21.

Milanovic, B. 2006. An estimate of average income and inequality in Byzantium around year 1000. *Review of Income and Wealth* 52(3): 449–470.

Milanovic, B., P.H. Lindert, and J.G. Williamson 2009. *Preindustrial inequality*. Unpublished ms, available at: http://econ.worldbank.org/projects/inequality. Previous version published as Measuring ancient inequality, National Bureau of Economic Research Working Paper No. 13550.

Morrisson, C., and W. Snyder. 2000. The income inequality of France in historical perspective. *European Review of Economic History* 4: 59–83.

Pareto, V. 1896. La courbe de la repartition de la richesse. Université de Lausanne. Republished as On the distribution of wealth and income, *Rivista di Politica Economica*, 645–660, 1997. (The same issue of *Rivista di Politica Economica* contains English translations of five other articles by Pareto on the same topic.)

Pen, J. 1971. *Income distribution: Facts, theories, policies*. New York/Washington, DC: Praeger.

Polanyi, K. 1944. *The great transformation*. Boston: Beacon Press.

Pomerantz, K. 2000. *The great divergence: China, Europe, and the making of the modern world economy*. Princeton: Princeton University Press.

Prados de la Escosura, L. 2007. Inequality and poverty in Latin America: A long-run exploration. In *The new comparative economic history: Essays in honor of Jeffrey G. Williamson*, ed. T.J. Hutton, K.H. O'Rourke, and A.M. Taylor, pp. 291–315. Boston: MIT Press.

Prados de la Escosura, L. 2008. Inequality, poverty, and the Kuznets curve in Spain, 1850–2000. *European Review of Economic History* 12: 287–324.

Rawls, J. 1971. *A theory of justice*. Cambridge, MA: Harvard University Press.

Rostovtzeff, M. 1926. *The social and economic history of the Roman Empire*, 1957. Oxford: Oxford University Press.

Scheidel, W., and S.J. Friesen 2009. *The size of the economy and the distribution of income in the Roman Empire*. Princeton/Stanford Working Papers in Classics, January.

Schiavone, A. 1995. *The end of the past: Ancient Rome and the modern West*, 2000. Cambridge, MA: Harvard University Press.

Schwarz, L.D. 1979. Income distribution and social structure in London in the late eighteenth century. *Economic History Review* 32: 250–259.

Shiue, C., and W. Keller 2007. Markets in China and Europe on the eve of the industrial revolution. *American Economic Review* 97: 1189–1216, September.

Soltow, L. 1968. Long-run changes in British income inequality. *Economic History Review* 21: 17–29.

Soltow, L. 1989. Income and wealth inequality in Amsterdam, 1585–1805. *Economisch-en-Social Historisch* 58: 72–95.

Soltow, L., and J.-L. van Zanden. 1998. *Income and wealth inequality in the Netherlands, 16th–20th century*. Amsterdam: Het Spinhuis.

Sussman, N. 2005. *Income inequality in Paris at the heyday of the commercial revolution*. Jerusalem: Hebrew University. Available at: http://economics.huji.ac.il/facultye/sussman/sussman_gobaleconomies.pdf.

Van Zanden, J.L. 1995. Tracing the beginning of the Kuznets curve: Western Europe during the early modern period. *Economic History Review* 48: 643–664.

Van Zanden, J.L. 2003. Rich and poor before the industrial revolution: A comparison between Java and the Netherlands at the beginning of the 19th century. *Explorations in Economic History* 40(1): 1–23.

Walbank, F.W. 1946. *The decline of the Roman Empire in the West*. London: Cobbett Press.

Wen, J.G. 2009. Why was China trapped in an agrarian society – An economicgeographical approach to the Needham puzzle. *New History*, forthcoming.

Williamson, J.G. 1980. Earnings inequality in nineteenth century England. *Journal of Economic History* 40: 457–476.

Williamson, J.G. 1982. The structure of pay in Britain, 1710–1911. *Research in Economic History* 7: 1–54.

Williamson, J.G. 1985. *Did British capitalism breed inequality?* Boston: Allen and Unwin.

Williamson, J.G. 2009. History without evidence: Latin American inequality since 1491. Mimeo.

Preobrazhensky, Evgenii Alexeyevich (1886–1937)

Michael Ellman

Abstract

Preobrazhensky was an Old Bolshevik and an original and perceptive Marxist theorist. His main contribution to Marxist political economy concerned the building of socialism in a predominantly agrarian country at a low level of economic development. He argued that socialist accumulation in such a country would require an initial period of original socialist accumulation. That is, economic growth on the basis of investment generated within industry would have to be preceded, in backward Russia with its limited industry, by a period of economic growth on the basis of investment resources obtained from outside the state sector.

Keywords

Accumulation of capital; Agriculture and economic development; Law of socialist accumulation; Marx's economics; New Economic Policy (NEP) (USSR); Over-accumulation; Preobrazhensky, E.; Socialism; Stalin, J.; Trotsky, L.

JEL Classifications

B31

An Old Bolshevik and a distinguished Marxist theoretician, Evgenii Alexeyevich Preobrazhensky joined the Russian Social Democratic Workers' Party (which split into Bolshevik and Menshevik factions) in 1903 and became a professional revolutionary, being repeatedly arrested and twice subject to internal exile. He led the local party organization in the Urals during the October Revolution. In 1918 he was a member of the Left Communist group within the party which opposed the treaty of Brest-Litovsk (which ended the Russian–German war by an agreement with 'imperialist' Germany rather than by a revolution within Germany). He played an active role in the Civil War (1918–20). He was a full member of the Central Committee of the Russian Communist Party (Bolsheviks) and also Central Committee Secretary in 1920–1. In 1921–2 he was critical of the New Economic Policy (NEP – a mixed-economy policy which permitted peasant households to utilize freely the land they cultivated and also permitted small-scale private enterprise in both villages and towns, while at the same time reserving the railways, large-scale industry, banking and international trade for the state). He was worried about concessions to the peasantry and their implications for rural stratification and Soviet power. A signatory to the Platform of the 46 (October 1923), he was an active oppositionist in 1924–7; he was expelled from the party in December 1927 and exiled to Siberia. Under the influence of Stalin's move to the Left, he broke with the Opposition and in July 1929 accepted Stalin's leadership. He attended the Seventeenth Party Congress (1934) where he praised Stalin and collectivization, denounced both himself and Trotsky (Stalin's chief political opponent), and advocated unity and unconditional acceptance of the party line and Stalin's leadership. Arrested in 1935, he served as a prosecution witness at the trial of Zinoviev (the former Politburo member and former chair of the executive committee of the Communist International) in 1936. Arrested again in 1936, he was not brought to a public trial, probably because of his refusal to confess to non-existent crimes. He was shot in 1937. In 1988 he was rehabilitated.

Preobrazhensky was the author of a large number of books and articles. They covered the exposition of Marxist-Leninist theory, financial and monetary questions, economic policy in France and economic policy in the USSR. Preobrazhensky's most original and important work concerned the problem of building socialism in a backward, overwhelmingly agrarian country.

Marx and Engels did not analyse how a future socialist economy would be organized and strongly opposed utopian socialism with its speculations divorced from current reality. Nevertheless, from their criticism of the anarchy of production under capitalism and their analysis of the views of rivals in the socialist movement, it is possible to draw inferences about how they expected a socialist economy to function. At the end of the 19th century Marxists had worked out some preliminary ideas for the transition to socialism and the organization of a socialist economy, as can be seen, for example, from the 1891 Erfurt Programme of the German Social Democratic Party and Kautsky's *Das Erfurter Programm* (1892), which is a commentary on it. They assumed, however, that the country concerned would be predominantly working-class and have a highly developed industry. In the 1920s, however, the Bolsheviks found themselves in power in a predominantly agrarian country at a low level of economic development. How should they build socialism in these circumstances? It is in answering this question that Preobrazhensky made his main contribution.

In *Novaia ekonomika* (1926a) he argued that, just as capitalist accumulation had required an earlier period of original accumulation as analysed in Marx (1867, vol. 1, part 8), so socialist accumulation would require an initial phase of original *socialist* accumulation. That is, economic growth on the basis of investment generated within industry would have to be preceded, in backward Russia with its limited industrial apparatus, by a period of economic growth on the basis of investment resources obtained from outside the state sector. He generalized his argument into a fundamental law of socialist accumulation which runs as follows:

> The more backward economically, petty-bourgeois, peasant, a particular country is which has gone over to the socialist organization of production, and the smaller the inheritance received by the socialist accumulation fund of the proletariat of this country when the social revolution takes place, by so much the more, in proportion, will socialist accumulation be obliged to rely on alienating part of the surplus product of pre-socialist forms of economy and the smaller will be the relative weight of accumulation on its own production basis, that is the less will it be nourished by the surplus product of the workers of socialist industry. Conversely, the more developed economically and industrially a country is, in which the social revolution triumphs, and the greater the material inheritance, in the form of highly developed industry and capitalistically organized agriculture, which the proletariat of this country receives from the bourgeoisie on nationalization, by so much the smaller will be the relative weight of pre-capitalist forms in the particular country; and the greater the need for the proletariat of this country to reduce non-equivalent exchange of its products for the products of the former colonies, by so much the more will the centre of gravity of socialist accumulation shift to the production basis of the socialist forms, that is, the more will it rely on the surplus product of its own industry and its own agriculture. (1926a, 1965 translation, p. 124)

As methods to obtain investment resources from the non-state sector (predominantly peasant agriculture), Preobrazhensky recommended the state monopoly of foreign trade, price policy, railway tariffs, taxation and state control of the banking system. He paid particular attention to the advantages of price policy as opposed to the use of coercion.

Preobrazhensky's analysis was very controversial when it was first published and led to a very heated debate. The reason for this is that the political basis of the Soviet regime in the 1920s was the precarious compromise between the Bolsheviks and the peasantry represented by the NEP. In addition, economic policy was based on the encouragement by the Bolsheviks for the peasants to 'enrich yourselves'. It was hoped that the development of peasant agriculture, in a mixed economy in which the commanding heights were in the hands of the state, would provide the food, raw materials, exports, internal market and labour force necessary for Soviet economic development. Hence Preobrazhensky's argument, with its presentation of the case for accumulation at the expense of peasant agriculture, was both politically and economically very disturbing. In particular, the analogy with original capitalist accumulation was distinctly ominous. According to Marx, original capitalist accumulation was based mainly on force, in particular on the use of force to expropriate the land from the peasantry. In the minds of the supporters of NEP, Preobrazhensky's analysis raised the spectre of a revival of the methods of War Communism (that is, requisitioning based on direct coercion, rationing, and attempted state control of the whole economy, rather than market economy methods).

Preobrazhensky's ideas evolved over time. In a paper of 1921 (1980, pp. 3–19), the very year the NEP was introduced, he anticipated an armed conflict between the Soviet state and the kulaks. He regarded this as inevitable and argued in good Stalinist style that 'the outcome of the struggle will depend largely on the degree of organization of the two extreme poles, but especially on the strength of the state apparatus of the proletarian dictatorship'. He concluded his argument, which was published at a time of serious famine and disease, partly caused by the class-war policies of the Bolsheviks, by warning his readers 'to prepare for everything that will ensure victory in the inevitable class battles that are to come'. In a paper of 1924, the thesis about the inevitable conflict between the state and the peasantry still plays a central role, but economic levers (for example, price policy) rather than coercion play the key role in resolving the conflict in the interests of socialist accumulation.

In a paper of 1927, attention has shifted to the conditions for growth equilibrium. The Harrodian conclusion about the essential precariousness of dynamic equilibrium is reached. The lesson is drawn that 'The sum of these contradictions shows how closely our development towards socialism is connected with the necessity – for not only political but also for economic reasons – to make a break in our socialist isolation and to rely in the future on the material resources of other socialist countries.'

In an unpublished paper of 1931 he criticized over-investment and pointed out the danger of an 'overaccumulation crisis'. His argument that 'socialism is production for consumption's sake' was unacceptable during the frenzy of the Soviet Great Leap Forward and was condemned as heretical. His position in 1931 seems to have been similar to that of Rakovsky, another Left Communist intellectual, who in an article of 1930 (published in 1931 and translated into English in 1981) warned against the coming Soviet economic crisis (which shook the whole economy in 1931–3) and stressed the wasteful and inefficient methods of Stalinist industrialization.

The accumulation that Preobrazhensky theorized about was *socialist* accumulation, that is, accumulation leading to the development of socialist relations of production. It is entirely natural, for example, that the imaginary author of Preobrazhensky's book *From NEP to Socialism* (1922), which takes the form of lectures supposedly given in 1970, is simultaneously a university professor and a fitter in a railway workshop. This reflected Preobrazhensky's expectation that the division of labour would be sharply reduced under socialism.

Preobrazhensky's work has had an enormous influence throughout the world. In the USSR in the 1920s he played a major role in the debate about the main directions of economic policy. In the West he was rediscovered in Erlich's famous paper in the *Quarterly Journal of Economics* (1950) and has been much discussed ever since. In the Third World his ideas play an important role in theoretical discussions and policy debates. He is rightly considered one of the outstanding Marxist economists of the 20th century.

See Also

- ▶ Agriculture and Economic Development
- ▶ Development Economics
- ▶ Marx's Analysis of Capitalist Production

Selected Works

1920. (With N. Bukharin.) *Azbuka kommunizma.* Petrograd. Trans. E. and C. Paul as *The ABC of communism.* Ann Arbor: University of Michigan Press, 1966.

1921. *Bumazhnye den'gyi v epokhu proletarskoi dictatury* [Paper money in the epoch of the proletarian dictatorship]. Tbilisi: Gosizdat.

1922. *Ot nepa k sotzializmu.* Moscow. Trans. B. Pearce as *From NEP to socialism.* London: New Park Publishers, 1973.

1924. *Ekonomicheskie krizisy pri NEP'e* [Economic crises under NEP]. Moscow: Izdatel'stvo Sotsialisticheskoi Akademii.

1926a. *Novaia ekonomika.* Moscow. Trans. B. Pearce as *The new economics.*Oxford: Clarendon Press, 1965.

1926b. *Ekonomika i finansy sovremennoi frantsii* [The economics and finances of contemporary France]. Moscow: Izdatel'stvo Kommunisticheskoi Akademii.

1927. Khozyaistvennoe ravnovesie v sisteme SSSR [Economic equilibrium in the system of the USSR]. *Vestnik kommunisticheskoi akademii,* No. 22.

1930. *Teoriia padaiushchei valiuty* [The theory of a depreciating currency]. Moscow: Gosizdat.

1931. *Zakat kapitalizma.* Moscow. Trans. R. Day as *The decline of capitalism.* New York: M.E. Sharpe, 1985.

1980. *The crisis of Soviet industrialization,* ed. D. Filtzer. London: Macmillan; New York: M.E. Sharpe. (This book of selected articles contains on pp. 237–240 a select bibliography of Preobrazhensky's works).

Bibliography

Day, R. 1975. Preobrazhensky and the theory of the transition period. *Soviet Studies* 27 (2): 196–219.

Erlich, A. 1950. Preobrazhenski and the economics of Soviet industrialization. *Quarterly Journal of Economics* 64 (1): 57–88.

Erlich, A. 1960. *The Soviet industrialization debate*. Cambridge, MA: Harvard University Press.

Filtzer, D. 1978. Preobrazhensky and the problem of the Soviet transition. *Critique* 9: 63–84.

Karshenas, M. 1995. *Industrialization and the agricultural surplus: A comparative study of economic development in Asia*. Oxford: Oxford University Press.

Kautsky, K. 1892. *Das Erfurter programm*. Stuttgart: Dietz.

Marx, K. 1867. *Das Kapital*. Vol. 1. Hamburg: Otto Meissner.

Millar, J. 1978. A note on primitive accumulation in Marx and Preobrazhensky. *Soviet Studies* 30 (3): 384–393.

Rakovsky, C. 1931. Na s"ezde i v strane [At the congress and in the country]. *Byulleten' Oppozitsii* (25–26): 9–32. Trans. D. Filtzer as 'The five year plan in crisis'. *Critique* (1981), No. 13.

Sah, R., and J. Stiglitz. 1984. The economics of price scissors. *American Economic Review* 74: 125–138.

Sah, R., and J. Stiglitz. 1987. Price scissors and the structure of the economy. *Quarterly Journal of Economics* 102 (1): 109–134.

Sah, R., and J. Stiglitz. 1992. *Peasants versus city-dwellers: Taxation and the burden of economic development*. Oxford: Oxford University Press.

Preordering

Charles Blackorby

Keywords

Difference principle (J. Rawls); Preordering; Reflexivity; Transitivity; Utilitarianism

JEL Classifications

D0

A preordering (also called a weak ordering or a quasi-ordering) is a reflexive and transitive binary relation which is not necessarily complete.

A binary relation R defined on a set S is a set of ordered pairs of elements of S, that is, a subset of the Cartesian product of S with itself, $S \times S$. One writes xRy (or $(x,y) \in R$) to mean that $x \in S$ stands in realtion R to $y \in S$. A preordering is a binary relation, R, which satisfies two properties: (i) reflexivity: for all $x \in SxRx$, and (ii) transitivity: for $x, y, z \in S$, if xRy and yRz, then xRz.

A simple example is given by the binary relation weak vector dominance which we denote V. Suppose S is Euclidean N-space, then xVy if and only if $x_n \geq y_n$, $n = 1,..., N$. V is clearly reflexive and transitive; it is just as clearly not complete, that is, not all elements of S are ranked. For example if $N = 2$, $x = (1, 2)$, and $y = (2, 1)$ then it is not the case that xVy or that yVx.

Quasi-orderings have played their largest role in welfare economics where consistency in decision making is a desirable requirement but where one may be dubious about being able to rank all possible outcomes. Two examples follow for which the notion of a subrelation is useful. Suppose R and S are binary relations: S is a subrelation of R if xSy implies xRy. For example, strong vector dominance, $\overline{V}$ is the binary relation which results when the above weak inequality is replaced with a strict inequality. Clearly, $\overline{V}$ is a subrelation of V.

Interpreting the elements of N-space as vectors of utilities, it is possible to define a quasi-ordering which is a subrelation of both the utilitarian and the Rawls criteria: Define the binary relation M by xMy

if and only if $\sum_{i=1}^{N} x_i \geq \sum_{i=1}^{N} y_i$ and $\min\{x_1, \ldots, x_N\} \geq \min\{y_1, \ldots, y_N\}$. M is clearly reflexive, transitive and not complete. The distributional insensitivity of the utilitarian principle is tempered by the Rawls's difference principle.

As an alternative, consider evaluating social states by weighted utility sums where the weights represent utility comparisons but these comparisons are not precisely fixed. Instead, the weights are drawn from a subset of N-dimensional Euclidean space, say B. More formally define the quasi-ordering F by xFy if and only if $\sum_{i=1}^{N} b_i x_i \geq \sum_{i=1}^{N} b_i y_i$ for all $(b_1, \ldots, b_N) \in B$. Suppose that we try to evaluate the desirability of burning down Rome while Nero fiddles. The quasi-ordering F may show a gain for burning Rome only if the set of interpersonal weights is such that Nero is given extreme consideration. (These examples are taken from the articles listed below.)

See Also

- ▶ Lexicographic Orderings
- ▶ Orderings
- ▶ Transitivity

Bibliography

Blackorby, C., and D. Donaldson. 1977. Utility vs. equity: Some plausible quasi-orderings. *Journal of Public Economics* 7: 365–382.

Sen, A.K. 1970. Interpersonal aggregation and partial comparability. *Econometrica* 38: 393–409.

Prescott, Edward Christian (Born 1940)

Stephen L. Parente

Abstract

\Edward Prescott was awarded the Nobel Prize in Economics in 2004 with Finn Kydland for their contributions to dynamic macroeconomics. Prescott is a member of a small group of economists who, starting in the 1970s, revolutionized macroeconomics by challenging the Keynesian consensus. While he is best known for his research on business cycles and the optimal design of economic policy, he has made important contributions to other applied fields, such as finance and development, as well as to economic theory. He has also made important contributions to methodology, having pioneered many of the standard contemporary techniques and tools in macroeconomics.

Keywords

American Academy of Arts and Sciences; American Economic Association; Bayesian analysis; Business cycle; Clubs; Consumer surplus; Debreu, G; Discretion vs rules; Dynamic programming; Econometric Society; Economic development; Endogenous growth models; Equity premium; Financial economics; Firms; Friedman, M; Gibrat's law; Haavelmo, T; Hodrick–Prescott filter; Imperfect information; Industrial organization; International inequality; Investment equation; Klein, L; Lotteries; Lucas, R; Microeconomic foundations; Misperceptions theory; Model calibration vs. statistical estimation; Modigliani, F; Moral hazard; Muth, J; Neoclassical growth model; New Classical Economics; Organizational capital; Partial equilibrium; Prescott, E; Productivity shocks; Rational expectations theory; Real business cycle theory; Recursive competitive equilibrium; Recursive methods; Relative incomes vs relative growth rates; Rules vs discretion; Statistical estimation vs. model calibration; Stochastic process; Time inconsistency; Total factor productivity

JEL Classifications

B31; B4; C6; C11; D8; E2; E3; E61; G1; L2; O4

Edward Christian Prescott is one of the leading macroeconomists of our time. He received the Noble Prize in Economics in 2004, an award he shared with Finn Kydland. Prescott's influence on the evolution of macroeconomics has been profound and far-reaching.

Prescott is a member of a small group of economists who, starting in the 1970s, revolutionized macroeconomics by challenging the Keynesian consensus that had held sway for half a century. This revolution, known as the ‘Rational Expectations Revolution’, occurred primarily at Carnegie Mellon University, the University of Chicago, the University of Minnesota, the University of Pennsylvania and Rochester University. Initially, this revolution was seen as the start of a new school of economic thought – New Classical Economics – that was based on the assumptions of market clearing and rational expectations. Today this revolution is seen simply as the start of an alternative approach to macroeconomics, one that advocates dynamic general equilibrium models with strong microeconomic foundations. This approach now dominates macroeconomics.

Prescott is best known for his applied research on business cycles, economic development and growth, and financial markets. In addition to his applied work, Prescott has produced a number of theoretical papers. A major line of this research demonstrates how to apply classical competitive analysis to economies with frictions, economies that previously had been regarded as outside the realm of such analysis. Within the profession, however, some of Prescott’s most lasting impacts have come in the methodology used for macroeconomic research. Many of the standard contemporary tools and techniques in macroeconomics were pioneered by Prescott. Finn Kydland and Edward Prescott together introduced calibrated models to macroeconomics; in doing so, they fundamentally changed the way applied macroeconomics is carried out. Prescott’s work has also been important for the development and diffusion of dynamic general equilibrium techniques and recursive methods. His work has also altered the ways in which economists handle data; for example, the so-called Hodrick–Prescott filter has become a standard tool of those working with time series data displaying trends. Thus, through applied work, theory, and methodology, Prescott has made lasting contributions to economics.

History

Edward C. Prescott was born in Glenn Falls, New York on 26 December 1940. He graduated from Glenn Falls High School in 1958 and Swarthmore College in 1962, with a BA in mathematics. In 1963 he received a Masters degree in Operational Research from Case University, which later became Case Western University. Thereafter, he enrolled in the Ph.D. programme at the Graduate School of Industrial Administration at Carnegie Mellon University (CMU), completing his doctorate in 1967.

Prescott started his academic career in 1967 as an assistant professor in the economics department at the University of Pennsylvania. In 1971 he left and accepted a position at CMU at the rank of assistant professor. He was promoted to the level of associate professor in 1972 and full professor in 1975. In 1974 he visited the Norwegian School of Business and Economics for a year at the invitation of Finn Kydland, who had written his dissertation under Prescott’s supervision at CMU. The visit is significant, as it was the occasion for much of the work for which the pair were later awarded the Nobel Prize. Prescott officially left CMU in 1980. Between 1978 and 1982, Prescott was a visiting professor at both the University of Chicago and Northwestern University. In 1981 he accepted a position at the University of Minnesota, where (with the exception of 1998, when he was a professor at the University of Chicago) he remained until 2003. At Minnesota Prescott was appointed a Regent’s Professor in 1996 and the McKnight Presidential Chair in Economics in 2003. In 2003 he left Minnesota to become the W.P. Carey Professor at the W.P. Carey Business School at Arizona State University. In addition to these academic positions, Prescott has served as an advisor to the Federal Reserve Bank of Minneapolis since 1981.

Prescott has received numerous honours in his distinguished career, with the Nobel Prize in 2004 being the most prestigious. He was elected to the Econometric Society in 1980 and the American Academy of Arts and Sciences in 1992. He is the recipient of the 2002 Erwin Plein Nemmers Prize in Economics, awarded by Northwestern

University biannually to an outstanding economist. He was chosen to give the First Lionel McKenzie Lecture at the University of Rochester in 1990, the Third Walras–Pareto Lecture at the Univeristé de Lausanne in 1994, and the First Lawrence Klein Lecture at the University of Pennsylvania in 1997. In addition, he was chosen to give the Richard T. Ely Lecture to the American Economic Association in 2002.

Research

Prescott's graduate training was largely in statistics. His thesis advisor was Mike Lovell. In addition, Maurice de Groot, one of the greatest Bayesian statisticians, was involved in the supervision of Prescott's work. Prescott's dissertation, titled 'Adaptive decision rules for macroeconomic planning', was an exercise in Bayesian statistical decision theory.

By his own admission Prescott was more of a statistician than an economist in the first years of his career. This changed in 1969, the year, he wrote 'Investment under uncertainty' with Robert E. Lucas (1971), whom Prescott had met while he was a graduate student at CMU. The paper studied the optimal investment decision of firms in an industry faced with stochastic demand. The paper was still in the tradition of the Keynesian 'system of equations' approach pioneered by Lawrence Klein that dominated macroeconomics at that time; its purpose was to derive a better investment equation to be used in large macro models. As such, it followed the trend established by Milton Friedman, Franco Modigliani, James Tobin and Trygve Haavelmo, who sought to base the individual equations in these systems on microeconomic theory.

The paper marks a watershed in the development of macroeconomics, however. It is one of the first to incorporate John Muth's hypothesis of rational expectations. The assumption of rational expectations forced Lucas and Prescott to develop and apply a new set of tools and concepts that have since become standard in macroeconomic research. For example, the paper introduced the concept of an equilibrium as a stochastic process. The paper is also important in the development of dynamic general equilibrium analysis; although the paper studied a partial equilibrium problem, the rational expectations assumption required that Lucas and Prescott simultaneously study the optimization problems of agents as well as the industry equilibrium. The paper also demonstrated how the competitive equilibrium could be solved from the social planner's problem of maximizing consumer surplus.

The 'Investment under uncertainty' paper is also important in that it showed how dynamic programming techniques developed in statistics and operations research could be successfully applied by economists to solve complicated optimization problems. In this respect, Prescott's previous training at Case University proved extremely valuable. The paper was to some extent a precursor to the concept of a 'recursive competitive equilibrium', that is, a set of time-invariant decision rules and prices that are functions of limited number of state variables.

In the 1970s Prescott continued to develop the recursive methods that now are commonplace in macroeconomics. In 1980, he published a paper with Rajish Mehra that extended and generalized recursive equilibrium theory. He also collaborated with Nancy Stokey and Robert Lucas in writing a comprehensive and complete book on the subject, *Recursive Methods in Economic Dynamics.*

Classical Competitive Analysis

The problems that Lucas and Prescott encountered in the 'Investment under uncertainty' paper necessitated that they read a number of papers in mathematical economics. One of these papers was Debreu's classic (1954) paper on competitive equilibrium analysis, 'Valuation equilibrium and Pareto Optimum'. This paper had a great impact on Prescott's thinking. First, the paper taught Prescott that many apparent market failures disappeared if mutually beneficial trades were permitted. Additionally, the paper showed to Prescott the power and importance of framing economic problems using the correct mathematical and verbal language.

Since the 1980s Prescott has written several papers that show how classical competitive analysis can be applied to a number of economies with frictions once the commodity space (that is, the set of tradable objects) is appropriately reframed. For many economies, the appropriate commodity space is defined over lotteries, namely, contracts with random components over goods and actions. Prescott and Townsend (1984) apply this approach to economies with moral hazard. Prescott and Rios-Rull (1992) do this in economies where people move between locations or occupations, and where at any one location there is imperfect information on the state of the other occupations. Hornstein and Prescott (1993) demonstrate how a number of potentially important production structures can also be mapped into this structure. Finally, Cole and Prescott (1997) show how this can be done in a class of economies where agents voluntarily form associations or clubs that carry out joint activities. These theoretical contributions were all intended to allow macroeconomists to address applied issues of policy relevance.

Industrial Organization

Prescott stopped teaching macroeconomics for a period in the 1970s, saying that it made little sense to teach a subject that one did not understand. During this period he primarily taught graduate courses in industrial organization (IO). A number of IO papers grew out of this teaching, such as Prescott (1973) and Prescott and Visscher (1977, 1980). Prescott and Visscher (1980) is an extremely important work. The paper shows that the acquisition of information, or organizational capital, by firms acts as an important cost of adjustment that limits firm growth. The paper makes an important contribution to the IO literature because it explains a number of empirical regularities, including Gibrat's law that firm size and growth are independent. It also makes an important contribution to the macroeconomic literature because the concept of organizational capital, which the paper introduced, has been used by a large number of researchers to understand a variety of phenomena.

Rules and Real Business Cycle Theory

During this period Prescott did not abandon research in the field of macroeconomics; in fact, he wrote with Finn Kydland two of the most important papers in macroeconomics: 'Rules rather than discretion: the time inconsistency of optimal plans', published in 1977, and 'Time to build and aggregate fluctuations,' published in 1982. The Nobel Prize Committee pointed to these two papers as the basis for awarding Kydland and Prescott the Noble Prize in 2004.

The substance of those papers is well known. Almost every undergraduate macroeconomic textbook written since the mid-1980s provides extensive coverage of both topics. The 'Time inconsistency' paper showed that people were made better off if the policymaker were to use a good rule instead of his discretion. Discretion – the ability of the policymaker to change his mind – leads to a worse outcome because the announced policy is typically not the optimal one to follow at the date of implementation. The 'Time to build' paper showed that productivity shocks account for roughly two-thirds of the volatility of US output over the business cycle in the post-war period. This productivity-driven view of the business cycle has come to be known as 'real business cycle theory'.

The idea that productivity shocks, which correspond to changes in the economy's stock of knowledge as well as changes in regulation or institutional factors, account for most of the US business cycle initially met fierce resistance. This was not surprising as it challenged the Keynesian view that the business cycle was a demand-driven phenomenon that called for government intervention on account of frictions. Many people objected to the theory on the grounds that the model contained no monetary side. These critics not only missed the point but they missed the fact that the precursor of the 'Time to build' paper (Kydland and Prescott 1980) did contain a monetary side based on Lucas's (1972) misperceptions theory. As that paper found that monetary shocks were quantitatively unimportant for understanding the US business cycle, Kydland and Prescott abstracted from money in their 1982 paper.

Over the years, resistance to Kydland and Prescott's theory of the business cycle has waned. The theory has been examined intensively, in fact probably more so than any other theory in economics to date. Attempts to discredit it have proven unsuccessful. Today, the idea that most of the US business cycle is a supply side phenomenon driven by productivity shocks is almost universally accepted.

The contribution of the 'Time to build' paper to the field of macroeconomics goes well beyond this substantive point. It also makes an important methodological point. Specifically, it lays the foundation for the use of deductive inference or model calibration to the application of macroeconomic issues. In this approach, the model is viewed as a measuring device, a so-called thermometer, to deduce or derive the quantitative implications of theory. This is in contrast to the inductive inference approach, or statistical estimation, that dominated the Keynesian 'system of equations' approach. There, the researcher attempts to discover the model out of a class of models that is the one to have most likely generated the data.

In effect, the 'Time to build' paper was an attempt to derive the quantitative implications of neoclassical growth theory for business cycles. Kydland and Prescott posed the question of whether the widely studied neoclassical growth model could be used not only to analyse long-run growth in the US economy but also to study business cycles. Starting with a standard neoclassical model, Kydland and Prescott restricted the values of the model parameters so that it quantitatively matched the trend growth of the US economy. They then modified the model so that productivity did not grow mechanically from year to year but instead followed a stochastic process that was based on the properties of 'Solow residuals' calculated from the data. Kydland and Prescott then reinserted these stochastic productivity shocks into the model and computed the equilibrium properties of the model. A striking result was that the model economy displayed business cycles that mirrored those found in the macro data, provided the labour supply was reasonably elastic.

Model calibration has become the dominant methodology in macroeconomics. It is widely used to test and develop theory as well as to evaluate policy. It is particularly useful in evaluating policy scenarios that are far 'out of sample' compared with historical experience. By using calibrated models, economists can conduct experiments on entire economies in a way that is not generally possible (or desirable!) with real economies.

Since 1982 Prescott and Kydland have continued to develop this line of research. They have subsequently written a number of articles to educate the profession in the use of model calibration (Kydland and Prescott 1991a, 1996; Prescott 2001). They have also written a number of articles that modify the model in their 1982 paper in order to explore further the implications of growth theory for understanding business cycles (Kydland and Prescott 1988, 1991b; Cooley et al. 1995).

The Equity Premium and International Income Differences

The application of the calibration methodology to a large number of macroeconomic questions has yielded important insights. As its founder, no person has used this methodology more effectively than Prescott. In addition to business cycles, Prescott's work has produced important insights in finance and economic development and growth.

Prescott's paper, 'The equity premium: a puzzle', co-authored with Rajnish Mehra and published in 1985, is a seminal work in financial economics. The paper sought to determine how much of the 6.2 percentage point difference between the average historical after-tax real rate of return on equity and the average historical after-tax return on bonds in the United States could be attributed to a premium for bearing non-diversifiable aggregate risk. The paper shows convincingly that households' aversion to risk cannot account for most of the difference in real rates of returns between bonds and equity. Prescott and Mehra's work has spawned an entire literature in financial economics, the goal of which is to solve the puzzle they uncovered.

Prescott's research has also fundamentally changed the way we think about the wealth and poverty of nations, and has set the direction of subsequent research in the area of economic development and growth. Prescott is among the very first researchers to argue that a theory of relative income levels, rather than relative growth rates, is the goal of development economics. Prescott did not start out with this view. Prescott's first paper on economic growth co-authored with John Boyd (Boyd and Prescott 1987) is an endogenous growth model whereby differences in policies or preferences across countries generate permanent differences in growth rates. After examining the development and growth facts over the period 1950–85 (Parente and Prescott 1993), however, Prescott concluded that the data did not support such a theory. The switch to a theory of relative income levels is evident in Parente and Prescott (1994). Today, the vast majority of papers that attempt to explain the huge disparity in international incomes take this relative income approach. Prescott is also one of the first researchers to have argued rigorously that differences in total factor productivity (TFP; that is, the efficiency with which a country uses its resources) account for most of the differences in international incomes. This is the main message of his 1997 Lawrence R. Klein Lecture, 'Needed: a theory of total factor productivity', published in 1998. Today, this view is almost universally accepted.

It should be no surprise that Prescott himself went on to provide a theory of TFP. In *Barriers to Riches*, Parente and Prescott (2000) demonstrate how a country's TFP is determined by policies that effectively constrain firms in their choice and use of technologies. Parente and Prescott (1999) completed the theory by arguing that these constraints typically exist to protect groups who stand to lose through the introduction of better technology.

No Decrease in TFP

Prescott has continued to remain highly productive; if anything, his productivity has increased in recent years. This recent research is mostly applied in nature. Like much of Prescott's previous work, it derives the implications of neoclassical growth theory for a variety of macroeconomic phenomena. One branch of this recent research uses the growth model to examine several longstanding hypothesis and puzzles in financial economics, including the equity premium (McGrattan and Prescott 2000, 2003, 2004). A different branch of this recent work uses the growth model to understand the reasons for the different experiences of OECD countries in the postwar period (Prescott 2003, 2004; Hayashi and Prescott 2002).

Teaching and Mentoring

It would be a serious omission not to mention Prescott's long-lasting commitment to teaching and advising. Prescott has supervised over 55 dissertations in his career, and many of his students are well-known economists such as Tom Cooley, Costas Aziaridis, Finn Kydland, Charlie Holt, Ed Green, Rajnish Mehra, Barbara Spencer, V.V. Chari, Hugo Hopenhayn, Richard Rogerson, Gary Hansen, Rody Manuelli, Gerhard Glomm, Jim Schmitz, Ayse Imrohoglu, Andreas Hornstein, Victor Rios- Rull, Fernando Alvarez, Dirk Krueger and Betsy Caucutt. Prescott's popularity as mentor reflects his philosophy that students should be treated as colleagues and that a good teacher has as much to learn from his students as they have to learn from him. The success of his students clearly speaks for the rigour that Prescott demands as well as the independence and confidence he instils in them. Prescott's students feel extremely fortunate to have had such a generous, engaging and inspiring advisor.

Conclusion

Edward C. Prescott is one of the most influential economists in the history of macroeconomics. His work has fundamentally changed the way economists conduct macroeconomic research and altered the way economists think about a large number of macroeconomic issues. Perhaps Robert E. Lucas in his introduction to Prescott's 2002 Richard T. Ely lecture best summarized Prescott's effect on economics, when he wrote

> We can remember the way we thought about the issues before Prescott's analysis, and the comparison with the way we think about them now gives a measure of the enormous effect each paper has had on our thinking... [Prescott's] papers met with resistance, but in the end [he has] caused us to rearrange important parts of our vision of how the economy works, to start over in many respects.

See Also

- ▶ Calibration
- ▶ Dynamic Programming
- ▶ General Equilibrium
- ▶ Growth and Cycles
- ▶ Inequality Between Nations
- ▶ Rational Expectations Models, Estimation of
- ▶ Recursive Competitive Equilibrium
- ▶ Total Factor Productivity

Selected Works

1971. (With R. Lucas.) Investment under uncertainty. *Econometrica* 39: 659–681.

1973. Market structure and monopoly profits: A dynamic theory. *Journal of Economic Theory* 6: 546–557.

1977. (With F. Kydland.) Rules rather than discretion: The time inconsistency of optimal plans. *Journal of Political Economy* 85: 473–491.

1977. (With M. Visscher.) Sequential location among firms with foresight. *Bell Journal of Economics* 8: 378–393.

1980. (With R. Mehra.) Recursive competitive equilibrium. *Econometrica* 48: 1365–1379.

1980. (With M. Visscher.) Organizational capital. *Journal of Political Economy* 88: 445–461.

1982. (With F. Kydland.) Time to build and aggregate fluctuations. *Econometrica* 50: 1345–1370.

1984. (With R. Townsend.) Pareto optima and competitive equilibria with adverse selection and moral hazard. *Econometrica* 52: 21–45.

1985. (With R. Mehra.) The equity premium: a puzzle. *Journal of Monetary Economics* 15: 145–161.

1987. (With J. Boyd.) Dynamic coalitions, growth and the firm. *American Economic Review* 67: 63–67.

1988. (With F. Kydland.) The work week of capital and its cyclical implications. *Journal of Monetary Economics* 21: 343–360.

1989. (With R. Lucas.) *Recursive methods economic dynamics.* Cambridge, MA: Harvard University Press.

1990. (With F. Kydland.) The workweek of capital and its cyclical implications. *Journal of Monetary Economics* 21: 340–360.

1991a. (With F. Kydland.) The econometrics of the applied general equilibrium approach. *Scandinavian Journal of Economics* 93: 161–178.

1991b. (With F. Kydland.) Hours and employment variations in business cycle theory. *Economic Theory* 1: 63–92.

1992. (With J. Rios-Rull.) Classical competitive analysis with islands. *Journal of Economic Theory* 57: 73–98.

1993. (With A. Hornstein.) The plant and the firm in general equilibrium theory. In *General equilibrium, growth, and trade II: The legacy of Lionel McKenzie,* ed. R. Becker et al. San Diego/Sydney/Toronto: Harcourt Brace.

1993. (With S. Parente.) Changes in the wealth of nations. *Federal Reserve Bank of Minneapolis Quarterly Review* 17: 3–16.

1994. (With S. Parente.) Barriers to technology adoption and development. *Journal of Political Economy* 102: 298–321.

1995. (With T. Cooley and G. Hansen.) An equilibrium analysis of idle resources and varying capacity utilization rates. *Economic Theory* 6: 35–50.

1996. (With F. Kydland.) The computational experiment: An econometric tool. *Journal of Economic Perspectives* 10: 68–86.

1997. (With H. Cole.) Valuation equilibrium with clubs. *Journal of Economic Theory* 74: 19–39.

1998. Needed: A theory of total factor productivity. *International Economic Review* 39: 525–552.

1999. (With S. Parente.) Monopoly rights: A barrier to riches. *American Economic Review* 89: 1216–1233.

2000. (With E. McGrattan.) Is the stock market overvalued? *Federal Reserve Bank of Minneapolis Quarterly Review* 24: 20–40.

2000. (With S. Parente.) *Barriers to riches.* Cambridge, MA: MIT Press.

2001. Business cycle theory: Methods and problems. In *Cycles, growth and structural change: theories and empirical evidence,* ed. L.P. Punzo. London: Routledge.

2002. (With F. Hayashi.) The 1990s in Japan: A lost decade. *Review of Economic Dynamics* 5: 206–235.

2003. (With E. McGrattan.) Average debt and equity returns: Puzzling? *American Economic Review* 93: 392–397.

2003. Prosperity and depression. *American Economic Review* 92: 1–15.

2004. Why do Americans work so much more than Europeans? *Federal Reserve Bank of Minneapolis Quarterly Review* 28: 2–13.

2004. (With E. McGrattan.) The stock market crash of 1929: Irving Fisher was right! *International Economic Review* 45: 991–1009.

Bibliography

Debreu, G. 1954. Valuation equilibrium and Pareto Optimum. *Proceedings of the National Academy of Sciences* 40: 588–592.

Lucas, R. 1972. Expectations and the neutrality of money. *Journal of Economic Theory* 4: 103–124.

Present Value

Stephen F. LeRoy

Abstract

The present value relation says that, under certainty, the value of a capital good or financial asset equals the summed discounted value of the stream of revenues which that asset generates. Otherwise arbitrage would be possible. Under uncertainty, and if risk neutrality is assumed, the future payoffs are replaced by their conditional expectations. Under risk aversion either the natural probability measure under which expectations are taken must be replaced by a 'risk-neutral measure', or the discount factor must be modified by a factor that reflects risk. The present value relation leads to bubbles if a convergence condition is not satisfied.

Keywords

Arbitrage; Bubbles; Capital asset pricing model; Capital budgeting; Capital market efficiency; Excess volatility tests; Fisher's separation principle; Martingales; Present value; Risk aversion; Risk neutrality; Risk premium; Risk-neutral probabilities; Speculative bubbles; Uncertainty; Wealth-maximization decision rule

JEL Classifications

G1

The present value relation says that, under certainty, the value of a capital good or financial asset equals the summed discounted value of the stream of revenues which that asset generates. The discount factor is that determined by the interest rate over the relevant period. The justification for the present value relation lies in the fact that (in perfect capital markets) an asset must earn a rate of return exactly equal to the interest rate.

Otherwise arbitrage opportunities emerge, which is inconsistent with equilibrium.

Derivation of the Present Value Relation

If r_t is the one-period interest rate at t, p_t is the (ex-dividend) price of an asset and d_t is its dividend, it must be true that

$$r_t = \left(d_{t+1} + p_{t+1}\right)/p_t - 1, \tag{1}$$

since the right-hand side equals the rate of return on the asset. Solving for p_t,

$$p_t = \frac{d_{t+1} + p_{t+1}}{1 + r_t}. \tag{2}$$

Replacing t by $t + 1$, (2) becomes an equation expressing p_{t+1} as a function of r_{t+1}, d_{t+1} and p_{t+2}. If the resulting expression is substituted to eliminate p_{t+1} in (2), and if this operation is repeated n times, it follows that

$$p_t = \sum_{i=1}^{n} \frac{d_{t+i}}{\prod_{j=0}^{i-1}\left(1 + r_{t+j}\right)} + \frac{p_{t+n}}{\prod_{j=0}^{i-1}\left(1 + r_{t+j}\right)}. \tag{3}$$

If speculative price bubbles are assumed not to occur (see below), the rightmost term in (3) converges to zero as n goes to infinity, so there results the present value equation

$$p_t = \sum_{i=1}^{\infty} \frac{d_{t+i}}{\prod_{j=0}^{i-1}\left(1 + r_{t+j}\right)}. \tag{4}$$

If in addition the interest rate is constant at $r_t = \rho$, (3) may be written as

$$p_t = \sum_{i=1}^{n} (1 + \rho)^{-i} d_{t+i} + (1 + \rho)^{-n} p_{t+n}. \tag{5}$$

or, if the convergence condition is satisfied, as

$$p_t = \sum_{i=1}^{\infty} (1 + \rho)^{-i} d_{t+i}. \tag{6}$$

In the special cases in which d_{t+i} is constant at d, or grows from d_t at rate g, (6) simplifies to

$$p_t = \frac{d}{\rho} \tag{7}$$

or

$$p_t = \frac{d_t(1 + g)}{\rho - g}, \tag{8}$$

respectively.

Present Value in Capital Budgeting

In introductory finance courses, the present value relation makes an early appearance in the chapter on capital budgeting, where it is taught that corporations should accept any investment project that promises a positive present value (net of costs), and only these. This wealth-maximization decision rule is the correct one independent of agents' preferences because, regardless of preferences, the consumption set that it generates dominates that generated by any other capital budgeting criterion. This is Irving Fisher's separation principle. Other criteria, such as accepting that project with the shortest payback period, or that with the highest internal rate of return, are either equivalent to present value maximization, ambiguous (sometimes, for example, a single project may have no real internal rate of return, or more than one) or wrong, depending on the characteristics of the project's returns.

Present Value Under Uncertainty

Under uncertainty, but on the assumption of risk neutrality, the present value relation may be written as

$$p_t = \sum_{i=1}^{\infty} (1 + \rho)^{-i} E_t(d_{t+i}), \tag{9}$$

which differs from (6) only in that future dividends is replaced by its conditional expectation.

This version of the present value relation has received extensive study, especially in the early finance literature. It is easily shown to imply

$$E_t(r_t) = \rho, \tag{10}$$

saying that the conditional expectation of the rate of return on the asset equals a constant independent of the conditioning set (Samuelson 1965, 1970). Here r_t is defined in (1); use of the same notation for the interest rate above and the rate of return on any asset here reflects the fact that under certainty the return on any asset equals the interest rate. This strong restriction provides the basis for most empirical tests of what has been called 'capital market efficiency' (Fama 1970; LeRoy 1989): if (10) is true, no information publicly available at t should be correlated with the rate of return on the asset from t to t + 1. In this sense prices 'fully reflect' all publicly available information.

The present value relation may also be interpreted from the vantage point of its martingale implication: if the asset is priced according to (9), the value x_t of a mutual fund which holds the asset and reinvests all of its dividend income will follow a martingale with drift, defined by

$$E_t(x_{t+1}) = (1+\rho)x_t. \tag{11}$$

To see this, assume that the mutual fund holds h_t shares of the asset so that

$$x_t = h_t p_t \text{ and } x_{t+1} = h_{t+1} p_{t+1}. \tag{12}$$

When dividend income is reinvested, h_{t+1} is given by the value that solves

$$h_{t+1} p_{t+1} = h_t \left(p_{t+1} + d_{t+1}\right). \tag{13}$$

Then

$$\begin{aligned} E_t(x_{t+1}) &= E_t\left(h_{t+1} p_{t+1}\right) \\ &= h_t E_t\left(d_{t+1} + p_{t+1}\right) = x_t(1+\rho). \end{aligned} \tag{14}$$

Here we used (1) and (10). To see that the correction for dividends payout is needed, observe that (10) implies that

$$\rho = \frac{E_t(d_{t+1})}{p_t} + \frac{E_t\left(p_{t+1}\right)}{p_t} - 1, \tag{15}$$

so that changes in the expected dividend yield are always offset one-for-one by changes in the expected rate of capital gain. If p_t by itself were a martingale the expected rate of capital gain would be a constant, implying that p_t is a constant multiple of expected dividends. But this is not an implication of the present value relation (take dividends as given by a first-order autoregressive process, for example). Hence p_t by itself does not follow a martingale.

The present value–martingale model appears in many contexts in finance. If a futures price is assumed equal to the conditional expectation of the relevant spot price, then the futures price will follow a martingale (Samuelson 1965). If owners of an exhaustible resource like petroleum extract it at optimal rates, then in some settings the price of reserves will appreciate according to a martingale with drift equal to the interest rate. Finally, the expected present value relation has implications for the volatility of asset prices. Informally, the expected present value relation implies that stock prices are like a moving average of the dividend stream to which they give title. Since a moving average is smoother than its components, it follows that stock prices should show less volatility than dividends. Volatility tests along these lines were originally reported by Shiller (1981) and LeRoy and Porter (1981). A number of subsequent papers extended and criticized the finding of excess volatility.

Equation (10), which requires that the conditional expectation of the rate of return does not depend on the value taken on by the conditioning variables, is very restrictive. Unlike its certainty analogue (1), which reflects only the assumption of zero transactions costs, (10) constitutes a strong restriction on the equilibrium probability distribution of the endogenously determined stock prices – much stronger than anything implied by the idea of capital market efficiency alone. The question becomes: what restrictions on preferences and the production technology are needed to derive (10)? LeRoy (1973) showed that, if agents are risk neutral, then in an exchange economy (10) will be satisfied (see also LeRoy 1982, for discussion in a more general setting). The result is a consequence of the obvious fact that if agents are risk neutral they will ignore moments in the distribution of rates of return higher than the first. Under nonzero risk aversion, however, the conditional expected rate of return will contain a risk premium which generally depends on the realizations of the conditioning variables. Hence (10) will generally not be true. LeRoy (1973) and Lucas (1978) discussed a class of models in which the expected present value property fails except as a special case.

If the assumption of risk neutrality is relaxed, the valuation equations must be changed. This can be achieved either by modifying the characterization of expected cash flows or by respecifying the discount factor. Modifying the characterization of expected cash flows involves distorting (relative to natural probabilities) the probability measure used to take expectations so as to put greater (lesser) weight on states in which agents have high (low) marginal utility. Such distorted probabilities always exist in the finite case, and exist under weak assumptions generally. When these ‘risk-neutral probabilities’ are used to compute expectations, security prices equal expected payoffs discounted at the interest rate, just as under risk neutrality (hence the name).

Alternatively, one can retain the natural probabilities but adjust the discount factor to allow for risk aversion. Under the capital asset pricing model, the risk premium on any security or portfolio is proportional to a beta coefficient, which equals the regression coefficient of the security’s return on that of the market portfolio. The constant of proportionality is the risk premium on the market portfolio. The idea is that risk-averse agents require high expected returns on high-beta securities since such securities increase portfolio risk on the margin.

Rational Bubbles

To return to the certainty case, if the rate of return on an asset is constant at ρ but the convergence condition

$$\lim_{n\to\infty}(1+\rho)^{-n}p_{t+n}=0 \tag{16}$$

is not satisfied, then asset prices are characterized by a rational speculative bubble. For a nontechnical introduction to rational bubbles, see LeRoy (2004). The asset’s price is higher than the present value of the stream of dividends the asset is title to, but nonetheless investors are willing to hold the asset because its price is expected to rise in the future. The definition of speculative bubbles under uncertainty is analogous (whether speculative bubbles exist or not has nothing directly to do with uncertainty). If speculative bubbles can occur, the present value Eq. (6) must be generalized to

$$p_t=\sum_{i=1}^{\infty}(1+\rho)^{-i}d_{t+i}+\gamma(1+\rho)^t, \tag{17}$$

where γ is an arbitrary non-negative constant capturing the magnitude of the speculative bubble. Equation (17) is the class of solutions to the difference equation

$$\rho=\frac{(d_{t+1}+p_{t+1})}{p_t}-1, \tag{18}$$

where γ is the constant of integration ($\gamma \geq 0$ results from the requirement that asset prices be always non-negative, a consequence of free disposal). In the special case $\gamma = 0$ speculative bubbles are absent and the present value relation results.

Bubbles cannot occur on a security that has a payoff only at one date, such as a zero-coupon bond. By induction, the same is true of securities that have payoffs at a finite number of dates. Existence of a bubble on such assets would imply an arbitrage opportunity: investors could sell the security short and purchase claims for its payoff at a cost equal to the present value of those payoffs. In the case of securities with payoffs at an infinite number of dates, it may or may not be possible to rule out bubbles on theoretical grounds. A few of the many papers dealing with this question are Tirole (1985), Gilles and LeRoy (1992a, b), Santos and Woodford (1997) and Huang and Werner (2000).

See Also

- ▶ Bubbles
- ▶ Excess Volatility Tests
- ▶ Finance
- ▶ Martingales
- ▶ Speculative Bubbles

Bibliography

Fama, E. 1970. Efficient capital markets: A review of theory and empirical work. *Journal of Finance* 25: 383–417.

Gilles, C., and S. LeRoy. 1992a. Asset price bubbles. In *The new Palgrave dictionary of money and finance*, ed. J. Eatwell, M. Milgate, and P. Newman. London: Macmillan.

Gilles, C., and S. LeRoy. 1992b. Bubbles and charges. *International Economic Review* 33: 323–339.

Huang, K., and J. Werner. 2000. Asset price bubbles in Arrow–Debreu and sequential equilibrium. *Economic Theory* 15: 253–278.

LeRoy, S. 1973. Risk aversion and the martingale model of stock prices. *International Economic Review* 14: 436–446.

LeRoy, S. 1982. Expectations models: A survey of theory. *Journal of Finance* 37: 185–215.

LeRoy, S. 1989. Efficient capital markets and martingales. *Journal of Economic Literature* 27: 1583–1621.

LeRoy, S. 2004. Rational exuberance. *Journal of Economic Literature* 42: 783–804.

LeRoy, S., and R. Porter. 1981. The present value relation: Tests based on implied variance bounds. *Econometrica* 49: 555–574.

Lucas, R. 1978. Asset prices in an exchange economy. *Econometrica* 46: 1429–1445.

Samuelson, P. 1965. Proof that properly anticipated prices fluctuate randomly. *Industrial Management Review* 6: 41–49.

Samuelson, P. 1970. The fundamental approximation theorem of portfolio analysis in terms of mean, variances and higher moments. *Review of Economic Studies* 37: 537–541.

Santos, M., and M. Woodford. 1997. Rational asset pricing bubbles. *Econometrica* 65: 19–57.

Shiller, R. 1981. Do stock prices move too much to be justified by subsequent changes in dividends? *American Economic Review* 71: 421–436.

Tirole, J. 1985. Asset bubbles and overlapping generations. *Econometrica* 53: 1499–1528.

Present Value of the Past

Charles Wolf Jr.

One of the seminal ideas in economics is that future events have a present value, which is calculable through a private or social rate of discount. Nevertheless, present valuation is not of much help for solving some problems, and can lead to erroneous results in others. For these problems, it is as important to have a method for evaluating the past as the future.

The Heuristics of Past Valuation

Economists usually view the past as concluded, and inert. It is not. At issue is both the unreliability of memory, and the influence of present actions on what is remembered. To explain important classes of events, and predict others, requires a change in the usual perspective.

Three premises underlie the present valuation of the past (Wolf 1970). The first is that prior events enter as arguments in the utility of various individual and institutional actors. Individuals and organizations are not only forward-looking maximizers, but backward-looking maximizers (or minimizers, where regret is concerned).

The second premise is that the aggregation of prior events may be likened to a process that employs a backward-looking discount rate, or ‘decay rate’. Various empirical proxies exist for measuring decay rates.

Thus, the stream of prior events that affects utility is subject to attrition in recall, and this attrition is reflected in the decay rate. The decay rate may also be negative: events from the past may wax, rather than wane, in recollection.

The third premise is that the decay rate may be affected by present actions. For example, a current event may increase or decrease the vividness with which a prior event is recalled (corresponding, respectively, to a decrease or an increase in the decay rate). Revisiting a place or a person, or confronting a new situation containing familiar characteristics, can have this effect. Hence, actions taken in the present, which may increase utility with respect to arguments having a present or future subscript, may diminish utility with respect to arguments that have a past subscript, and vice versa.

Modeling the Present Value of the Past (PVP)

The process of present valuation of the past may be clarified by a simple structural model (Wolf 1970). (1) A utility function expressing current welfare, U_t, as a function of current income, Y_t, and the present value of a set of prior events, $\tilde{V}_t$. Y_t can also be viewed as the present value of future income discounted to the present in the usual manner, $\tilde{V}_t$ is often associated with status, reputation, and self-esteem. (The utility function is thus similar to games that combine status and welfare (Shubik 1975).) (2) A function that specifies the present value of prior events, $\tilde{V}_t$ in terms of their values at the time of their original occurrence, *V,* mediated to the present by the decay rate, r^*. (3) A decay rate function, expressing r^* as a function of a particular choice, *i,* among a set of *n* feasible policies, P_{ti}, for *changing* income in period *t*. (4) An income growth equation (associated with (3) above), which specifies $\dot{Y}_t$ as depending on the policy choice P_{it}. (5) An income identity defining current income as prior income Y_{t-1}, plus the income change, $\dot{Y}_t$.

Some Examples of PVP: Sunk Costs and Social Inequities.

Sunk Costs

According to a familiar economic theorem, sunk costs should not influence decision, only marginal costs. To the ordinary person's complaint ('prior costs really do matter'), the usual response is either an evasion ('people are just irrational'), or a tautology (marginal costs can be redefined to include pain associated with an otherwise preferred choice because of its connection with prior events). Present valuation of the past provides a more satisfactory explanation.

The sunk costs example applies to many situations in which present action is influenced by a desire to preserve a present benefit whose magnitude is indicated by the scale of prior (sunk) *costs*.

A classic example is provided by Agamemnon's stratagem for persuading the Greeks to preserve in the Trojan wars by pointing out that withdrawal would cause dishonour to those whose lives had already been lost (Homer, *Iliad*).

If two alternative actions have the same expected yield but the decision maker has previously expended resources on one, which will (should) he choose? The descriptive answer is easy; the normative answer involves more elusive considerations of prestige, credibility, and the desire for personal vindication, which can be readily related to PVP. When the yields are not equal, the proper analytic precept is not to choose the higher, but rather to show exactly how much higher it is, so that this margin, ΔY_t, can be compared with the possible loss of other (prior) values, $\Delta \tilde{V}_t$, in arriving at a utility-maximizing choice.

PVP also applies to a familiar phenomenon sometimes associated with ageing. People often find it tolerable, or even gratifying, to acknowledge their currently inadequate performance if they can thereby claim that it contrasts with their superior performance in prior years ('You should have seen me ten years ago!').

People seem to normalize for time in relation to their past, which they have accumulated in different amounts. One implication for the PVP model is that an increase in current income, *Yt,* may seem smaller in relation to changes in the decay rate, r^*, and to the present value of the past, $\tilde{V}_t$ the older a person is. People who have lived longer simply have more sunk costs tied up in PVP, and changes in the decay rate are therefore relatively more important to them than changes in current income.

The Negative Present Utility of Past Injustice

Over a century ago, de Tocqueville posed a well-known paradox based on his studies of the French Revolution (de Tocqueville 1856). De Tocqueville's paradox can be formulated in more general terms: Why is it that improvements in welfare and in social justice often intensify resentment and unrest?

In this case, PVP is the discounted aggregate of prior costs or injuries. It is likely to be a heavily weighted argument in the utility functions of these who believe they have been previously exploited.

As with sunk costs, the choice of a policy, P_{ti}, for influencing present income affects the present value of the past by changing the decay rate, r^*. If a rise in current income lowers the decay rate, the sense of past injustice becomes more vivid and painful, and the negative present utility of prior inequalities is thereby magnified.

Thus, it is not necessarily the largest positive change in current income, $\dot{Y}_t$ that will make the biggest positive contribution to utility. A lower $\dot{Y}_t$ may be preferable because it does not lower the decay rate (and hence, the present value of the past), as much. Moreover, there may be no current incomes policy which in fact raises utility. The choice may be between allowing current income to stagnate, or raising current income, but actually lowering utility (hence, resentment and violence) if utility is to be raised at some future time.

Bibliography

de Tocqueville, A. 1856. *The old régime and the French Revolution*. New York: Doubleday, 1955. Homer. *Iliad*. Trans. R. Lattimore. Chicago: University of Chicago Press, 1951.

Shubik, M. 1975. *Games for society, business and war: Towards a theory of gaming*. New York: Elsevier.

Wolf Jr., C. 1970. The present value of the past. *Journal of Political Economy* 78(4): 783–792.

Pretesting

Jan R. Magnus

Abstract

This article briefly discusses the meaning and dangers of pretesting in estimation procedures. It outlines the proof of the equivalence theorem, and compares the pretest estimator with three other estimators: the 'usual' estimator, the 'silly' estimator and the 'Laplace' estimator.

Keywords

Estimation; Model selection; Pretesting

JEL Classifications

C12; C13

Model Selection Versus Estimation

Suppose data are generated by a linear relationship

$$y = X\beta + \gamma z + u, \quad u \sim N\left(0, \sigma^2 I_n\right), \qquad (1)$$

where X is an $n \times k$ matrix of explanatory variables and z is an additional $n \times 1$ vector of explanatory variables. In our role as investigator we do not know this relationship. Our interest is in the effect of X on y, that is, we want to estimate β. Since we don't know that y is generated by (1), we formulate a model that will serve as a vehicle to estimate β. Let us assume that we know that the relationship is linear, that X is certainly in the model, and that z is perhaps in the model. For simplicity we assume also that σ^2 is known. Thus our 'model space' consists of only two models: the unrestricted model (where $\gamma \neq 0$) and the restricted model (where $\gamma = 0$).

Our interest could be in finding the 'true' model, in which case we are concerned with *model selection*. In that case we should select the unrestricted model, however small γ turns out to be. Our interest, however, is in the *estimation* of β and the model is not of interest *per se* – it is only a means towards our goal. Even if we knew that γ is nonzero, this would not necessarily mean that we should include z in our regression. This is because, if γ is close to zero, a small bias in the estimates of β will result if we use the restricted model, but their variances may increase substantially, and hence the mean squared error will also increase substantially. (Recall that the bias depends on the value of γ but the variance does not.) So even if we know the truth, it is typically wise to simplify for the purposes of estimation.

What Is Pretesting?

The ordinary least squares (OLS) estimator for β in the restricted model is of course

$$b_r = (X'X)^{-1}X'y. \tag{2}$$

If we define

$$M = I_n - X(X'X)^{-1}X', q = \frac{\sigma}{\sqrt{z'Mz}}(X'X)^{-1}X'z,$$
$$\theta = \frac{\gamma}{\sigma/\sqrt{z'Mz}},$$

then we can write the OLS estimators for β and γ in the unrestricted model as

$$b_u = b_r - \hat{\theta}q, \quad \hat{\gamma} = \frac{z'My}{z'Mz}, \tag{3}$$

where

$$\hat{\theta} = \frac{\hat{\gamma}}{\sigma/\sqrt{z'Mz}} = \frac{z'My}{\sigma\sqrt{z'Mz}} \sim N(\theta, 1) \tag{4}$$

denotes the t-ratio, which is normally distributed in this case because σ^2 is assumed known. We call θ the *theoretical* t-ratio.

Since we don't know which of the two models we should use in order to estimate β, the typical econometric practice is to perform a preliminary test (pretest) on γ, and to include z in our regression if the t-ratio $\hat{\theta}$ is 'large' and exclude it if $\hat{\theta}$ is 'small'. This leads to the so-called *pretest estimator*

$$b = \begin{cases} b_r & \text{if} \quad \left|\hat{\theta}\right| \leq c, \\ b_u & \text{if} \quad \left|\hat{\theta}\right| > c, \end{cases} \tag{5}$$

where c is some positive number such as 1.96. We can also write (5) as

$$b = \lambda b_u + (1-\lambda)b_r, \quad \lambda = \begin{cases} 0 & \text{if} \quad \left|\hat{\theta}\right| \leq c, \\ 1 & \text{if} \quad \left|\hat{\theta}\right| > c, \end{cases} \tag{6}$$

which emphasizes that the pretest estimator is a weighted average of the estimators in the available models. The weights, however, are random variables because they depend on $\hat{\theta}$ The pretest estimator is therefore a complicated nonlinear estimator.

The problem with pretesting is not so much that people do it, but that they ignore the consequences. In typical econometric practice, model selection takes place using t-ratios and other diagnostics, after which a single model is selected (stage 1). Then estimates and standard errors are obtained in the selected model (stage 2), and these are reported. It is then tacitly assumed that the reported estimates are unbiased and that their standard errors are given by the usual OLS formulae. This assumption, however, is incorrect. The estimates are biased and their standard errors are not given by the usual OLS formulae. This is the pretest problem.

The Equivalence Theorem

Things are made simpler by the equivalence theorem, originally proved by Magnus and Durbin (1999), and improved and extended by Danilov and Magnus (2004a).

Theorem 1 (Equivalence Theorem) Let $b = \lambda b_u + (1 - \lambda)b_r$, where $0 \leq \lambda \leq 1$ and $\lambda = \lambda(My)$.

Then, letting $\widetilde{\theta} = \lambda\hat{\theta}$, we have

$$E(b) = \beta - E\left(\widetilde{\theta} - \theta\right)q, \mathrm{var}(b) = \sigma^2(X'X)^{-1} + \mathrm{var}\left(\widetilde{\theta}\right)qq'$$

and hence

$$MSE(b) = \sigma^2(X'X)^{-1} + MSE\left(\widetilde{\theta}\right)qq'.$$

Proof

We know from (3) that $b_u = b_r - \hat{\theta}q$, so that

$$b = \lambda b_u + (1 - \lambda)b_r = b_r - \lambda\hat{\theta}q = b_r - \hat{\theta}q.$$

The crucial ingredient is that b_r and My are independent, so that

$$E(b_r|My) = E(b_r), \quad \mathrm{var}(b_r|My) = \mathrm{var}(b_r).$$

Also, since both λ (by assumption) and $\hat{\theta}$ as given in Eq. (4) depend only on My, we see that $\hat{\theta} = \lambda\hat{\theta}$ depends only on My. Hence,

$$E(b|My) = E(b_r) - E\left(\hat{\theta}|My\right)q = \beta + \theta q - \widetilde{\theta}q$$
$$= \beta - \left(\widetilde{\theta} - \theta\right)q$$

and

$$\mathrm{var}(b|My) = \mathrm{var}(b_r|My) = \mathrm{var}(b_r)$$
$$= \sigma^2(X'X)^{-1}.$$

Now using the well-known relationships between conditional and unconditional moments, we obtain

$$E(b) = E(E(b|My)) = \beta - E\left(\widetilde{\theta} - \theta\right)q,$$

and

$$\mathrm{var}(b) = E(\mathrm{var}(b|My)) + \mathrm{var}(E(b|My))$$
$$= \sigma^2(X'X)^{-1} + \mathrm{var}\left(\widetilde{\theta}\right)qq',$$

and hence

$$\begin{aligned} MSE(b) &= \mathrm{var}(b) + E(b - \beta)(b - \beta)' \\ &= \sigma^2(X'X)^{-1} + \mathrm{var}\left(\widetilde{\theta}\right)qq' \\ &\quad + \left(E\left(\widetilde{\theta} - \theta\right)\right)^2 qq' \\ &= \sigma^2(X'X)^{-1} + MSE\left(\widetilde{\theta}\right)qq'. \end{aligned}$$

This completes the proof. ||

The equivalence theorem is important because it tells us that if we have a 'good' estimator for θ, say $\widetilde{\theta}$, then this defines $\lambda = \widetilde{\theta}/\hat{\theta}$ and *the same* λ will provide a good estimator for β, namely $b = \lambda b_u + (1 - \lambda)b_r$. The pretest estimator chooses

$$\widetilde{\theta} = \begin{cases} 0 & \text{if} \quad \left|\hat{\theta}\right| \leq c, \\ \hat{\theta} & \text{if} \quad \left|\hat{\theta}\right| > c, \end{cases}$$

which is not a good choice as we shall see.

Moments of the Pretest Estimator

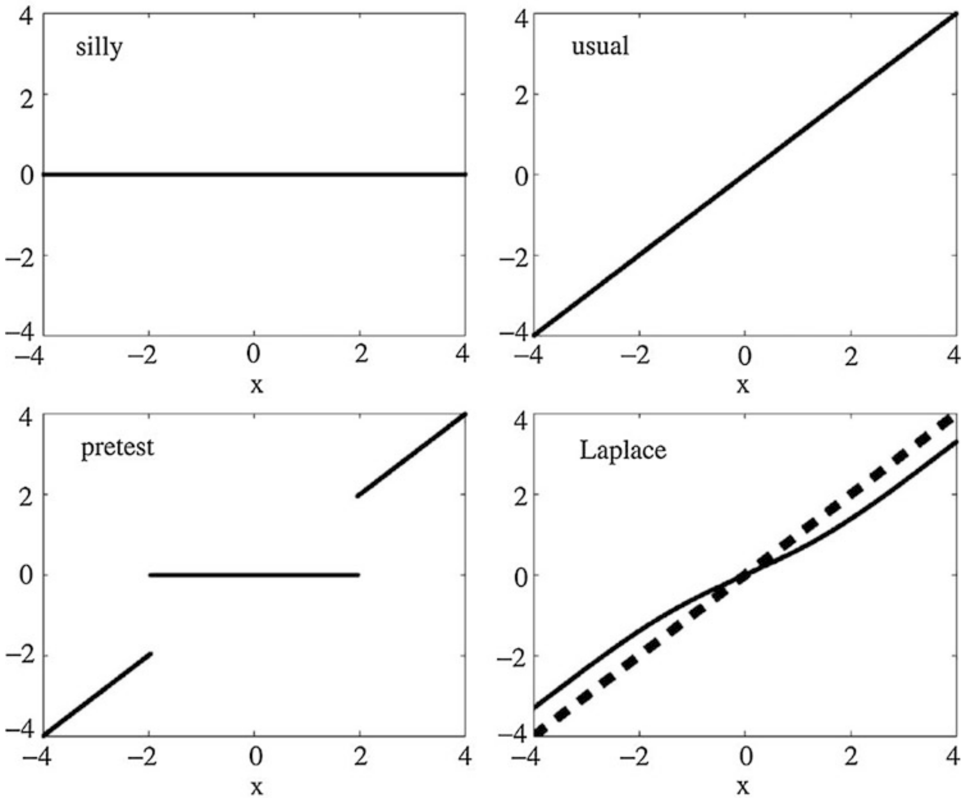
In the previous section we have seen that the pretest estimator is, in essence, of the form

$$t(x) = \begin{cases} 0 & \text{if} \quad |x| \leq c, \\ x & \text{if} \quad |x| > c, \end{cases} \tag{7}$$

where $x \sim \mathrm{N}(\theta, 1)$. When studying this estimator, we confront it with three other estimators: the 'usual' estimator $t(x) = x$, the 'silly' estimator $t(x) = 0$, and the 'Laplace' estimator introduced in Magnus (2002). The four estimators are graphed in Fig. 1 for $|x| < 4$.

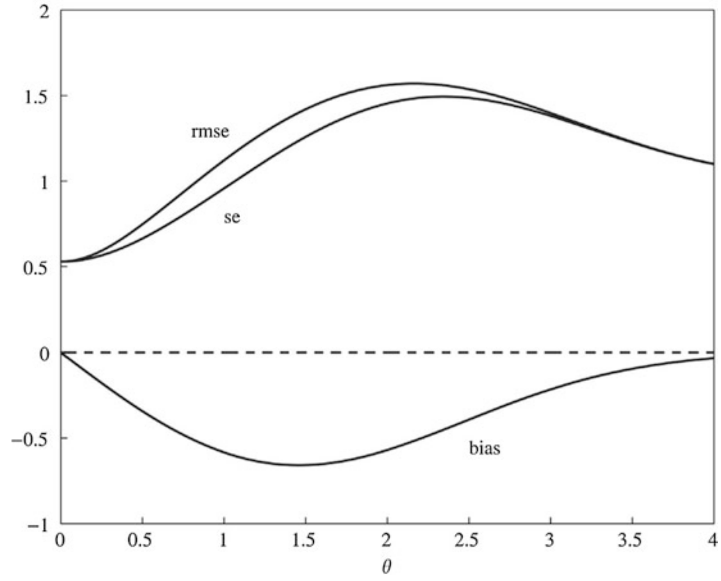
It is clear that the pretest estimator is discontinuous, hence inadmissible. But this is only one of its uncomfortable properties.

Theorem 2 (Moments of Pretest Estimator) Let $x \sim \mathrm{N}(\theta, 1)$ and let $t(x)$ be the pretest estimator defined in (7). Then,



Pretesting, Fig. 1 Four estimators $t(x)$ of θ

Pretesting, Fig. 2 Moments of the pretest estimator



$$E(t-\theta) = \phi(c-\theta) - \phi(c+\theta) - \theta P$$

and

$$E(t-\theta)^2 = 1 + (c+\theta)\phi(c+\theta) + (c-\theta)\phi(c-\theta) + \left(\theta^2 - 1\right)P,$$

where ϕ denotes the standard-normal density and $P = \int_{-\theta-c}^{-\theta+c} \phi(u)du.$

Proof
Letting $S = \{u : -\theta - c < u < -\theta + c\}$, we have

$$\begin{aligned} &E\,(t(x)) \\ &= \int_{-\infty}^{\infty} t(x)\phi(x-\theta)dx = \int_{|x|>c} x\phi(x-\theta)dx \\ &= \theta - \int_{|x|<c} x\phi(x-\theta)dx \\ &= \theta - \int_S (u+\theta)\phi(u)du \\ &= \theta - \int_S u\phi(u)du - \theta\int_S \phi(u)du \\ &= \theta + [\phi(u)]_S - \theta P = \theta + \phi(-\theta + c) \\ &-\phi(-\theta - c) - \theta P, \end{aligned}$$

using the fact that $\phi'(u) = -u\phi(u)$. Similarly, using the fact that $\phi''(u) = (u^2 - 1)\phi(u)$, we obtain the second result. ||

The bias, standard error and root mean squared error of the pretest estimator are graphed in Fig. 2.

We see that the bias is relatively small compared with the standard error. Since $bias(-\theta) = -\,bias(\theta)$ (so that θ and $bias(\theta)$ have opposite signs), and since we know from Theorem 1 that $bias(b_i) = -bias\left(\widetilde{\theta}\right)q_i$, we can determine the direction of the pretest bias.

Theorem 3 (Sign of Pretest Bias) Let $w := (X'X)^{-1}X'z$ with components w_i $(i = 1, \ldots, k)$. Then the pretest bias of b_i is positive (that is, $E(b_i) > \beta_i$) if and only if $\gamma_i w_i > 0$. As a consequence we can estimate the sign of the pretest bias of b_i by $sign(w_i\hat{\gamma}_i)$.

For purposes of exposition we have concentrated on the simplest case, but considerable generalization is possible to more than one additional z-variable, to unknown σ^2, and to general variance matrix.

Alternatives

We now compare the pretest estimator with the four estimators in Fig. 1. We graph the root mean squared error (RMSE) of each of the four estimators in Fig. 3.

The ‘usual’ estimator is unbiased and has variance one, independent of the value of θ. The ‘silly’ estimator is obviously better when θ is close to zero, the two estimators have the same RMSE when $\theta = 1$, corresponding to the fact that

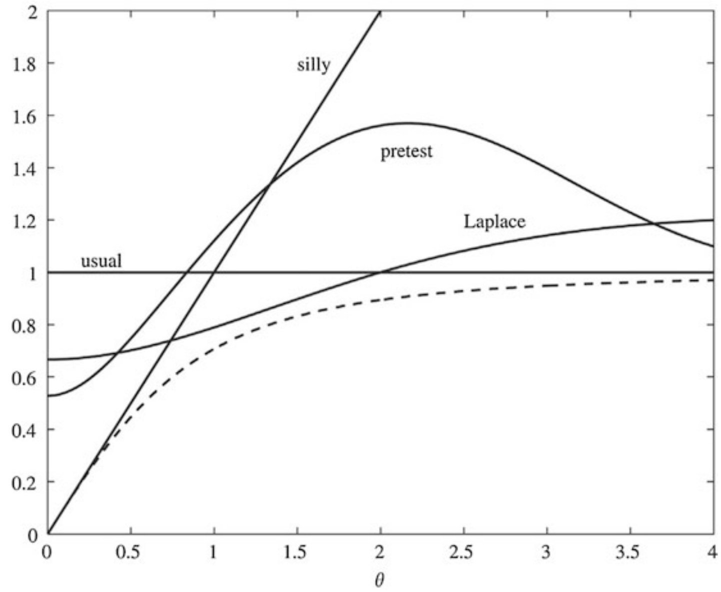
$$MSE(b_r) - MSE(b_u) = \left(\theta^2 - 1\right)qq',$$

but the RMSE of the ‘silly’ estimator is unbounded. The pretest estimator lies in-between the silly and the usual estimator, except in the important interval around $\theta = 1$ where the pretest estimator is *worse* rather than better than either of the two naive alternatives. This is a most unwelcome property of the pretest estimator, and it has given rise to thought about alternatives. An attractive alternative is the so-called Laplace estimator, which has a Bayesian and a non-Bayesian interpretation, is admissible, is based on a ‘neutral’ prior, and has good properties around $\theta = 1$. The dotted line $|\theta|/\sqrt{1+\theta^2}$ denotes the theoretical lower bound of the root mean squared error.

History

The implications of model selection on the estimators of the parameters of interest were already being discussed following Tinbergen’s (1939) study for the League of Nations. Both Keynes

Pretesting, Fig. 3 Root mean squared error of the four estimators



(1939) and Friedman (1940), in their respective critiques on Tinbergen, focused on the method of model selection when the estimation procedure repeatedly uses the same data to discriminate between plausible competing theories. The same point was made in Haavelmo (1944, Section 17). Koopmans (1949) suggested that a completely new theory of inference was required to solve the dilemmas implied by the model selection problem.

Early work on the pretest estimator includes Bancroft (1944, 1964), Huntsberger (1955), Larson and Bancroft (1963), Cohen (1965), Wallace and Ashtar (1972), Sclove, Morris and Radhakrishnan (1972), Bock, Yancey and Judge (1973b), and Bock, Judge and Yancey (1973a). The harm of ignoring the effects of pretesting was analysed by Danilov and Magnus (2004a, b). Important surveys are provided by Judge and Bock (1978, 1983), Judge and Yancey (1986), Giles and Giles (1993), and Magnus (1999).

See Also

- ▶ Model Selection
- ▶ Robust Estimators in Econometrics
- ▶ Semiparametric Estimation
- ▶ Testing

Acknowledgments I am grateful to Chris Müris for preparing the figures.

Bibliography

Bancroft, T.A. 1944. On biases in estimation due to the use of preliminary tests of significance. *Annals of Mathematical Statistics* 15: 190–204.

Bancroft, T.A. 1964. Analysis and inference for incompletely specified models involving the use of preliminary tests of significance. *Biometrics* 20: 427–442.

Bock, M.E., G.G. Judge, and T.A. Yancey. 1973a. Some comments on estimation in regression after preliminary tests of significance. *Journal of Econometrics* 1: 191–200.

Bock, M.E., T.A. Yancey, and G.G. Judge. 1973b. The statistical consequences of preliminary test estimators in regression. *Journal of the American Statistical Association* 68: 109–116.

Cohen, A. 1965. Estimates of linear combinations of the parameters in the mean vector of a multivariate distribution. *Annals of Mathematical Statistics* 36: 78–87.

Danilov, D., and J.R. Magnus. 2004a. On the harm that ignoring pretesting can cause. *Journal of Econometrics* 122: 27–46.

Danilov, D., and J.R. Magnus. 2004b. Forecast accuracy after pretesting with an application to the stock market. *Journal of Forecasting* 23: 251–274.

Friedman, M. 1940. Review of Jan Tinbergen, Statistical Testing of Business Cycle Theories, II: Business Cycles in the United States of America. *American Economic Review* 30: 657–661.

Giles, J.A., and D.E.A. Giles. 1993. Pre-test estimation and testing in econometrics: Recent developments. *Journal of Economic Surveys* 7: 145–197.

Haavelmo, T. 1944. The probability approach in econometrics. *Econometrica* 12(Suppl): 1–115.
Huntsberger, D.V. 1955. A generalization of a preliminary testing procedure for pooling data. *Annals of Mathematical Statistics* 26: 734–743.
Judge, G.G., and M.E. Bock. 1978. *The statistical implications of pre-test and Stein-Rule Estimators in econometrics*. Amsterdam: North-Holland.
Judge, G.G., and M. E. Bock. 1983. Biased estimation. In *Handbook of econometrics*, Vol. 1, ed. Griliches Z. and M. D. Intriligator. Amsterdam: North-Holland, Chapter 10.
Judge, G.G., and T.A. Yancey. 1986. *Improved methods of inference in econometrics*. Amsterdam: North-Holland.
Keynes, J.M. 1939. Professor Tinbergen's method. *Economic Journal* 49: 558–568.
Koopmans, T. 1949. Identification problems in economic model construction. *Econometrica* 17: 125–144.
Larson, H.J., and T.A. Bancroft. 1963. Biases in prediction by regression for certain incompletely specified models. *Biometrika* 50: 391–402.
Magnus, J.R. 1999. The traditional pretest estimator. *Theory of Probability and Its Applications* 44: 293–308.
Magnus, J.R. 2002. Estimation of the mean of a univariate normal distribution with known variance. *Econometrics Journal* 5: 225–236.
Magnus, J.R., and J. Durbin. 1999. Estimation of regression coefficients of interest when other regression coefficients are of no interest. *Econometrica* 67: 639–643.
Sclove, S.L., C. Morris, and R. Radhakrishnan. 1972. Non-optimality of preliminarytest estimators for the mean of a multivariate normal distribution. *Annals of Mathematical Statistics* 43: 1481–1490.
Tinbergen, J. 1939. *Statistical testing of business cycle theories*. Vol. 2. Geneva: League of Nations.
Wallace, T.D., and V.G. Ashtar. 1972. Sequential methods in model construction. *The Review of Economics and Statistics* 54: 172–178.

Price Control

John Kenneth Galbraith

Abstract

In modern times price control has been used to keep down food prices, as part of prices and incomes policies, in wartime economic management, to help governments win elections, and to tackle inflation. Along with macroeconomic restraint and specific commodity restraint by rationing, price controls used by the Allies in the Second World War succeeded in countering inflation. The ineffectiveness of price control in Latin America has helped give it a bad name. There are radically different forms of controls in greatly differing contexts; price control should be seen as a diversely applicable policy, sometimes advisable, sometimes not.

Keywords

Black market; Corporations; Craft guilds; Depressions; Fiscal policy; Galbraith, J. K.; Incomes and prices policy; Inflation; Latin America; Monetary policy; Price control; Rationing; Rent control; Second World War; Trade unions; Wage-price inflation

JEL Classifications

H0

The fixing of prices by public action is of exceedingly ancient origin; popular economic cliché associates it with the Edict of Diocletian, and economic history dwells at length on the controls exercised and imposed by the medieval guilds. Only in modern times, roughly the last 200 years, has it fallen under the general disapproval and interdict of orthodox economic attitudes and has it been seen therein as a temporary or aberrant departure from free-market principles.

In a more adequate view, controls have not one but several forms, some of which are, in their context, a reflection of necessary and appropriate policy, as other designs in other contexts are not.

Specifically, some five employments of price controls can be identified, apart from public-utility and like regulation which reflects the different and largely accepted need to maintain a public surveillance and restraint on natural or legislated monopoly power. There is:

(1) The use of controls to address particular wartime pressures of demand on supply, as in the

United States and other countries in World War II, and to keep down the price of food for urban dwellers, as now in African countries and elsewhere. These can perhaps be called episodic or casual controls.

(2) The use of controls as part of what has come to be called a prices and incomes policy. They here act on the specific problem of wage/price inflation.

(3) The use of controls as part of a comprehensive exercise in wartime economic management, backed by rationing of consumers' goods, allocation of materials and labour and a general restraint on aggregate demand.

(4) The use of controls as a highly temporary expedient to get by an election.

(5) The use of controls in the face of an enduring inflationary movement propelled by a persisting excess of aggregate demand, as recently in Latin America and of late in Israel.

Two of the above employments of controls – to limit the wage/price dynamic and as an adjunct to a comprehensive mobilization of economic resources, as in World War II – have modern policy relevance. In the highly organized modern economy of strong corporations and viable and effective trade unions, price inflation can come from the microeconomic effect of prices and living costs on wages and of wage demands on prices. Much recent experience shows that this wage/price dynamic can be arrested by conventional monetary or fiscal action only by the restraining force of substantial unemployment on wage demands and much idled plant capacity on prices. In other words, conventional monetary and fiscal policy arrest wage/price inflation only as they cause a recession or depression.

Accordingly, attention has focused on direct restraint by the state. Avoiding the unduly blunt reference to wage and price controls, this has come to be called 'an incomes and prices policy'. Austria, Germany, Japan and other industrial countries have, formally or informally, resorted extensively to such restraints. The English-speaking countries and their economists, businessmen and unions have been more reluctant.

Market forces must not be impaired. Still, by a growing minority such restraints are viewed as a necessary alternative to economically and politically more painful designs for restraining wage/ price inflation. There continue to be repetitive suggestions that such intervention distorts the market allocation of resources. Mention is not made of the way that strong unions and strong corporations in the modern economy have already invaded resource-allocation procedure and accommodated it to their purposes.

A more serious problem lies outside the field of economics. Where fiscal and monetary restraints need only a negligible administrative apparatus, any effective form of price and wage control requires a substantial administrative one. And, of course, the public intervention to limit price or wage increases is highly visible. Thus it invokes the ever-present suspicion or dislike of government intervention and bureaucracy. Fiscal and monetary action, even when more painful in overall effect, encounters far less resistance.

The second acceptable form of controls was the comprehensive design used in all of the industrial countries in World War II. In combination with macroeconomic restraint on demand by fiscal policy and specific commodity restraint by rationing, such controls successfully countered the threat of price inflation in Britain, the United States, Canada and other participants in those years. In the case of non- rationed non-essentials, the controls substituted shortage or non-availability for rationing by price. Some evasion of controls by black market operations was present, but this, though greatly publicized and deplored, was, in general, relatively limited.

Once controls were fully in place in the United States, price increases were nominal, and, overall, there is no memory of inflation from the war years. Increases then or following the removal of controls were insignificant as compared with the double-digit inflation, as it was called, of the 1970s.

The circumstances, especially in the United States during World War II, were, however, exceptionally and perhaps uniquely favourable for successful use of price controls. After ten years, depression had come to be considered by the early 1940s a normal and inevitable peacetime phenomenon. Accordingly, after the war, unemployment and associated hardships would recur. From this came a powerful incentive to save – to save, among other reasons, for the cars and others durables that would only later be available. Labour, in effect, was employed against the promise of future consumption. At the same time there was in the United States the large increase in the supply of non-durables as previously unemployed plant and labour were drawn into production. Overall civilian consumption increased, and this further reduced the pressure of demand on the controls. A similar *general* use of controls following a period of high employment and serious or even incipient inflation with associated expectations would be a far more difficult matter.

Of the other uses of controls there is less to be said. Isolated or piecemeal controls (in contrast with a broad-based incomes and prices policy) can, indeed, have the effect of diverting resources from the area of control and into uncontrolled and thus more remunerative employments. This has been a consequence of one of the more persistent manifestations of isolated or piecemeal controls, that of rents. Its yet more serious manifestation has been in the poor countries, notably of Africa. There the use of price controls to keep down food costs has been an important contributing cause of the food distress and famine.

Controls in the face of a massive excess of demand have been a frequent resort in Latin America. A case can be made for such action to alter expectations, themselves a cause of inflationary pressures of demand, before instituting strong macroeconomic restraints. More frequently, such controls have been a separate and often desperate response to demand-induced inflation. The resulting evasion, ineffectiveness and eventual collapse have contributed notably to the poor reputation of controls in general.

In 1971–3, President Richard Nixon used general controls with great effect to suppress wage–price inflation and allow of companion fiscal and monetary support to employment. Largely, if not principally, in consequence, he carried every state but Massachusetts and the District of Columbia in the election of 1972. Such success for controls must, however, be accounted for and judged in the field of politics, not economics. The removal of the controls after the election restored with some precision the circumstances that had led to their being involved with a strong recurrence of inflation.

A common tendency of orthodox economics has been to deal with all forms of price control as a homogeneous exercise. This, it will be evident, is a grave oversimplification. In fact, there are radically different forms of controls in greatly differing contexts. Reasonable and, indeed, necessary sophistication requires that these differences be recognized, that price control be seen as a diversely applicable policy, sometimes greatly advised, sometimes wholly the reverse.

See Also

► Command Economy

Bibliography

Galbraith, J.K. 1952. *A theory of price control*. Cambridge, MA: Harvard University Press.

Keynes, J.M. 1940. *How to pay for the war*. London/New York: Macmillan/Harcourt, Brace and World.

Mitchell, H. 1947. The edict of Diocletian: A study of price fixing in the Roman empire. *Canadian Journal of Economics and Political Science* 13: 1–12.

Rockoff, H. 1984. *Drastic measures: A history of wage and price controls in the United States*. Cambridge: Cambridge University Press.

Taussig, F.W. 1919. Price-fixing as seen by a price-fixer. *Quarterly Journal of Economics* 33 (2): 205–241.

Price Discrimination

Louis Phlips

Price discrimination is as common in the market place as it is rare in economics textbooks. It appears under many disguises and explains a large number of business practices which are difficult to rationalize otherwise. Its ubiquity results from the fact that there is price discrimination whenever two varieties of a commodity are sold (by the same seller) to two buyers at different *net* prices, the net price being the price (charged to the buyer) corrected for the cost associated with what differentiates one variety from another. Transportation and storage costs are examples that readily come to mind. Costs of product design and of services offered by distributors are less obvious examples (as is the cost associated with demand uncertainty). Given such costs, there is no price discrimination when these costs are fully reflected in the prices. Price discrimination typically implies that part of these costs is 'absorbed': delivered or future prices increase by less than the cost of carrying the good over space or time; models of better quality are sold at a better price–quality ratio; better service is not fully charged. Alternatively, a product or service produced at the same cost is offered at a price that decreases as the quantity bought increases.

This type of pricing is profit maximizing when the seller has some monopoly power, the opportunity to sell in several sub-markets and therefore the possibility of maximizing overall profits (rather than maximizing profits separately in each region, in each time period, or for each particular product specification). It works only to the extent that (a) arbitrage (i.e. transfers of commodities between sub–markets) is impossible or costly; (b) customers can be sorted according to the intensity of their demand; and (c) their demands are in fact different. When these conditions are met, a discriminating price policy is more profitable than a nondiscriminating one: if you can price discriminate, it is always profitable to do so. The reason is simple. Compared with a uniform price, discriminating prices are not only closer to the highest price a particular customer is ready to pay (his 'reservation price'): they also make it possible to serve customers who would not be able to buy at the uniform price or to induce them to consume more than they would otherwise.

Pigou (1920, pt. 2, ch. 17) makes a useful distinction between three types of price discrimination. *First-degree* discrimination 'would involve the charge of a different price against all the different units of commodity, in such wise that the price exacted for each was equal to the demand price for it, and no consumers' surplus was left to the buyers' (p. 279). *Second-degree* discrimination is an approximation to perfect discrimination. It obtains when a firm is able to make n separate prices such that all units with a reservation price greater than p_1 are sold at the price p_1, all with a reservation price less than p_1 and greater than p_2 at a price p_2, and so on.

Buyers are separated into n groups and there is one (identical) price per group, so that some buyers have a consumers' surplus. All buyers with a reservation price greater than p_n are served. *Third-degree* discrimination would obtain if the monopolist were able to distinguish among his customers n different groups, separated from one another more or less by some practicable mark, and could charge a separate monopoly price to the members of each group . . ..This degree, it will be noticed, differs fundamentally from either of the preceding degrees, in that it may involve the refusal to satisfy, in one market, demands represented by demand prices in excess of some of those which, in another market, are satisfied (p. 279).

While second-degree discrimination refers to individual reservation prices for one unit of a commodity, third-degree discrimination refers to sub-market demand curves. Let there be three adjacent spatial sub-markets with different demand curves and let \$90, \$95, and \$100 be the three profit maximizing delivered prices. Suppose the transportation cost per unit from one sub-market to the next (and back) is \$10 (so that there is freight absorption). This is third-degree price discrimination, since buyers located in the market with demand prices smaller than $100 but higher than $95 are unable to buy anywhere, the cost of transportation being prohibitive. The same is true for buyers located in the second market with demand prices lower than $95.

First- and Second-degree Price Discrimination

First-degree price discrimination is often, understandably, called ‘perfect’. On the one hand, it is the most profitable form of discrimination, since it extracts the entire consumer surplus. On the other hand, each buyer is able to buy the product he or she wants to buy and reveals the value of the product by paying the highest price he or she is ready to pay.

In real life, frequently used techniques are to bargain reductions in prices or in carrying costs. (One possible result is that higher rates may be charged for carrying passengers or freight over a short distance than over a long distance – see Friedman 1979.) Another commonly used practice is to offer a two-part tariff. Such tariffs are feasible when the product cannot be resold at a reasonable cost among consumers – a requirement met by most service industries. They are easy to implement when a connection or entrance fee (the first part) can be charged in addition to price per unit consumed (the second part).

The economics of two-part tariffs can best be understood with reference to the pricing policy of an amusement park such as Disneyland. In his seminal paper, Oi (1971) showed that the best way to extract the entire consumers’ surplus is to charge the highest possible entrance fee to each visitor and thus to discriminate at the entrance gate, with the implication that the price of a ride inside the amusement park should be set equal to its marginal cost – since the surplus cannot be extracted twice from the same visitor. Such a two-part tariff is Pareto optimal. In practice, it is approximated by charging *n* different entrance fees to *n* different groups of visitors: discounts are granted to senior citizens, groups, military, children, etc.

Block tariffs such as those charged by public utilities for the supply of electricity, gas, water, etc. often take the form of a series of two-part tariffs, the total expenditure on a first block being the entrance fee to the next (cheaper) block. Consumers are free to decide which block they want to be in (by adjusting their consumption) and thus reveal how much they are ready to pay.

In fact, each consumer ends up paying a different average price depending on the quantity consumed. Prices are quantity dependent or ‘non-linear’ and decrease with the quantity consumed. Defining ‘reservation outlays’ (the maximum amount a consumer in a given income class is ready to pay for a given number of units of the good) as increasing functions of income, Spence (1977, 1980) is able to show that a profit-maximizing seller will extract the entire surplus from the lowest income class and proportionally less as income increases (in order to persuade the richest customers to buy the largest quantities) but will provide the optimal quantity to the richest customers. When quantities are replaced by qualities, similar conclusions arise (Phlips 1983, ch. 14). On using reservation prices (per income class) for purchase in a particular time period, the same analysis leads to the conclusion that the richest customer, though first to buy a new product at the highest price, will keep the highest surplus while the poorest will be the last to be served with no surplus left. This is the so-called ‘skimming pricing’ (Dean 1949, pp. 419–21; Stokey 1979).

A move from a uniform price to a non-linear price can be welfare improving. Leland and Meyer (1976) and Littlechild (1975) have shown that such a move benefits consumers in the aggregate, with the gainers able to compensate those who lose. Willing (1978) and Spence (1980) show that for any price different from marginal cost, there is a change to a non-linear outlay schedule that directly benefits all consumers, without the need for transfers effected outside the market.

First and second-degree discrimination can also result from tying the sale of one good to that of another, where the first is typically a durable (a copying machine) and the second a nondurable (paper). Here it is profit-maximizing to lease the durable at a very low rental price (a very low entrance fee, in fact) to attract as many users as possible and to extract the individual surplus through a high price per copy made. When tying is impossible (e.g. because ordinary paper can be used on the copying machine), the intensity of use of the machine is measured with the help of a meter. By his or her decision on the number of copies made, each individual user reveals his or her reservation price *for the machine*! And this value is extracted via the price per copy. If the machine were sold at a uniform price, fewer consumers would be served and profits would be lower (Telser 1965, 1979). An even more subtle form of surplus extraction is to offer simultaneously two or more goods separately and in a bundle, at a special (lower) price for the bundle (Adams and Yellen 1976).

Third-degree Price Discrimination

A profit-maximizing monopolist facing *n* markets with different demand curves will set the *n* prices in such a way that marginal revenues are equal from market to market. Indeed, the *n* first order conditions form a system of equations, each of which equates the marginal revenue in a particular market to the marginal cost of production. And since this marginal cost takes a unique value (the value associated with total production), all marginal revenues must be equal to it and therefore to each other. This is the fundamental (third-degree) price discrimination rule.

In an abstract world, in which the causes of market separation are not specified, it is not possible to give a general answer to the question of whether third-degree price discrimination increases output or welfare. The discussion therefore concentrates on particular specifications of the demand curves and of the marginal cost curve. For example, Robinson (1933, bk 5) has shown geometrically – and Schmalensee (1981) has generalized this result – that if a single-price monopoly selling in two markets under constant costs is allowed to discriminate between them, total output is unchanged if both markets have linear demand curves. This result depends critically on the unrealistic assumption that markets are served under *both* regimes. When some form of product differentiation (spatial, temporal, etc.) is introduced, price discrimination typically serves to open *new* markets and thus to increase sales and welfare.

Consider a spatial monopolist serving several adjacent market areas. Marginal revenue equalization implies that only delivered prices are quoted and that part of the freight is absorbed. (If market area demands are linear, exactly one half of the cost of transportation between two points in space is absorbed (Beckmann 1976).) As a result, distant markets that would not be served under a uniform f. o.b. pricing policy can be reached.

Next consider intertemporal profit maximization by a monopolist that can produce for inventory. Commenting on Smithies' pioneering 1939 paper, Shaw (1940, p. 469) remarks:

> There is an evident parallelism between the theory of inventory accumulation and discriminating monopoly. The former defines optimum distribution of a total supply between markets that are separated in time. The latter defines optimum distribution of a total supply between markets that are separated in space. Assuming appropriate discounting of future quantities, the definitions of optima are identical . . ..

The equalization of discounted marginal revenues over time leads to price stickiness over time and normal cost pricing (Phlips 1983, ch. 6). In the special case where inventories are produced by nature in the form of a stock of exhaustible resources, intertemporal price discrimination gives Hotelling's 1931 rule: the monopoly price of an exhaustible resource increases at a rate that is smaller than the real rate of interest.

The study of Nash equilibria for markets in which oligopolists discriminate is a new area of intensive research. The emergence of more or less uniform delivered prices as the result of oligopolistic competition in space is one of the new insights provided by Greenhut and Greenhut (1975), Norman (1981) and Neven and Phlips (1985). (See also the symposium edited by Phlips and Thisse 1982.) Intertemporal price discrimination in a dynamic game is shown by Phlips and

Richard (1985) to imply stagflation when the real rate of interest is positive. These preliminary results suggest that the study of games in which oligopolists are allowed to price discriminate will lead to a better understanding of pricing as it occurs in the real world.

See Also

- ▶ Consumer Surplus
- ▶ Discriminating Monopoly
- ▶ Dumping
- ▶ Spatial Competition

Bibliography

Adams, W.J., and J. Yellen. 1976. Commodity bundling and the burden of monopoly. *Quarterly Journal of Economics* 90: 475–498.

Beckmann, M.J. 1976. Spatial price policies revisited. *Bell Journal of Economics* 7: 619–630.

Dean, J. 1949. *Managerial economics*. Englewood Cliffs: Prentice-Hall.

Friedman, D.D. 1979. In defense of the long-haul/short-haul discrimination. *Bell Journal of Economics* 10: 706–708.

Greenhut, J., and M.L. Greenhut. 1975. Spatial price discrimination, competition, and locational effects. *Economica* 42: 401–419.

Hotelling, H. 1931. The economics of exhaustible resources. *Journal of Political Economy* 39: 137–175.

Leland, H.E., and R. Meyer. 1976. Monopoly pricing structures with imperfect information. *Bell Journal of Economics* 7: 449–462.

Littlechild, S. 1975. Two-part tariffs and consumption externalities. *Bell Journal of Economics* 6: 661–670.

Neven, D., and L. Phlips. 1985. Discriminating oligopolists and common markets. *Journal of Industrial Economics* 34(2): 133–149.

Norman, G. 1981. Spatial competition and spatial price discrimination. *Review of Economic Studies* 48: 97–111.

Oi, W.Y. 1971. A Disneyland dilemma: Two-part tariffs for a Mickey Mouse monopoly. *Quarterly Journal of Economics* 85: 77–96.

Phlips, L. 1983. *The economics of price discrimination*. Cambridge: Cambridge University Press.

Phlips, L., and J.F. Richard. 1985. *A dynamic oligopoly model with demand inertia and inventories*, CORE discussion paper.

Phlips, L., and J.-F. Thisse, eds. 1982. Symposium on spatial competition and the theory of differentiated markets. *Journal of Industrial Economics* 31(1–2): September–December.

Pigou, A.C. 1920. *The economics of welfare,* 4th edn. London: Macmillan, 1922.

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

Schmalensee, R. 1981. Output and welfare implications of monopolistic third-degree price discrimination. *American Economic Review* 71: 242–247.

Shaw, E.S. 1940. Elements of a theory of inventory. *Journal of Political Economy* 48: 465–485.

Smithies, A. 1939. The maximization of profits over time with changing cost and demand functions. *Econometrica* 7: 312–318.

Spence, M. 1977. Nonlinear prices and welfare. *Journal of Public Economics* 8: 1–18.

Spence, M. 1980. Multi-product quantity-dependent prices and profitability constraints. *Review of Economic Studies* 47: 821–841.

Stokey, N.L. 1979. Intertemporal price discrimination. *Quarterly Journal of Economics* 93: 355–371.

Telser, L.G. 1965. Abusive trade practices: An economic analysis. *Law and Contemporary Problems* 30: 488–505.

Telser, L.G. 1979. A theory of monopoly of complementary goods. *Journal of Business* 52: 211–230.

Willig, R. 1978. Pareto-superior nonlinear outlay schedules. *Bell Journal of Economics* 9: 56–69.

Price Discrimination (Empirical Studies)

Frank Verboven

Abstract

Price discrimination occurs when the prices of similar products sold by the same firm show variation that cannot be attributed to cost variation. Recent empirical work has identified the presence of both direct and indirect price discrimination, after cost-based explanations have been accounted for. Furthermore, there is increasing evidence on the sources of price discrimination. The extent of price discrimination has often been found to increase as competition intensifies, in contrast to conventional wisdom but consistent with new theoretical insights. Finally, various empirical studies have considered the effects of price discrimination on profits, consumer welfare and efficiency.

Keywords

Competition and price discrimination; Cross-price elasticities; Gini coefficient; Monopoly price discrimination; Multi-product firms; Poaching; Price discrimination; Price elasticity; Single-product firms

JEL Classifications

L10

Price discrimination occurs when the prices of similar products sold by the same firm show variation that cannot be attributed to variation in marginal costs. Direct (or third-degree) price discrimination serves to exploit observed differences in consumer characteristics; indirect (or second-degree) price discrimination exploits unobservable consumer heterogeneity. While price discrimination has been studied extensively by economic theorists, and illustrated with numerous textbook examples (for example, Scherer and Ross 1990), it has only recently become an area of rigorous empirical research. Empirical studies have focused on several questions: (*a*) the measurement or identification of price discrimination; (*b*) the sources of price discrimination, notably the role of competition; and (*c*) the effects of price discrimination on profits, consumer welfare and efficiency.

Measurement of Price Discrimination

The identification of price discrimination can be introduced in a simple framework in which a firm sells two products. The price difference Δp between the two products (assumed positive) can be decomposed in a cost difference Δc and a margin difference Δm, so $\Delta p = \Delta c + \Delta m$. Price discrimination exists to the extent that the observed price difference Δp is due to the margin difference Δm rather than a possible cost difference Δc. (An alternative definition is based on percentage rather than absolute margin differences. To consider this, reinterpret the variables in logs, Clerides 2004, compares the two approaches in empirical studies.) Identifying margin differences from cost differences is not an obvious task. Lott and Roberts (1991) provide plausible cost-based explanations for commonly viewed price discrimination cases. Several recent empirical studies have attempted to properly account for cost differences before drawing conclusions about the presence of price discrimination.

There have been two methodological approaches. The first approach uses direct cost information. Sometimes the cost difference can be derived from industry information about the production technology. An early example is Benston (1964), who finds that 76 per cent of the higher interest rates charged to small businesses can be attributed to additional costs. In contrast, Clerides (2002) attributes only five per cent of the average price difference between hardback and softback books to higher production costs. In other cases, the production technology is not known, but the cost difference Δc is reasonably assumed to be zero or negative, so that the observed positive price difference Δp provides a lower bound for the extent of price discrimination. Graddy (1995) finds that Asians pay seven per cent less at a fish market, while there are no reasons to believe that these customers have lower servicing costs. Degryse and Ongena (2005) find that bank customers pay lower interest rates as their distance from the bank increases, whereas costs, if anything, are expected to be increasing in distance. Shepard (1991) provides a neat variation on this theme. As in the above framework, she observes the price difference Δp between a high-quality and a low-quality product sold by multi-product firms (full service and self-service at gas stations). In addition, she essentially also observes the analogue price difference Δp^S for single-product firms (selling either full-service or self-service). She defines the extent of price discrimination as the difference between the markup difference for multi-product firms Δm and that of single-product firms Δm^S. Because her qualitative evidence indicates that the cost difference Δc between the high-quality and low-quality product for multi-product firms is no larger than the corresponding cost difference Δc^S for single-product firms, the difference between Δp and Δp^S provides a lower bound for the extent of price discrimination. She finds that the extent of price discrimination for full-service versus self-service gasoline amounts to at least nine cents a gallon.

The second approach to identifying price discrimination does not use cost information, but instead infers the price–cost margin difference Δm from a model of pricing behaviour. This approach essentially replaces cost-side information by demand-side information. For example, Verboven (2002) finds evidence of indirect price discrimination between high-mileage and low-mileage drivers. He uses information on the relative popularity of high-quality and low-quality products (diesel and gasoline cars) and the distribution in consumers' willingness to pay for quality (mileage). He infers that 75–90 per cent of the price premium for the high-quality products can be attributed to a higher margin, a finding that is confirmed by direct cost information.

Sources of Price Discrimination

Several empirical studies have gone beyond the basic question of identifying price discrimination to uncover its sources, in particular the role of competition. Theoretical work has revealed that competition does not necessarily reduce the incentives to price discriminate. The extent of direct price discrimination depends on both the price elasticity of market demand and the cross-price elasticities with respect to competing products; it is therefore not necessarily smaller under competition. For example, Borenstein (1991) looks at price discrimination in the competitive retail gasoline market. Margins on unleaded gas were initially higher than margins on leaded gas. The decline in the number of competing stations offering leaded gas caused an increase in the margins on leaded gas relative to the margins on unleaded gas, hence a reduction in price discrimination. This illustrates that competition can be a source of price discrimination: stations take into account the buyers' possibilities to substitute to competing stations when setting their prices. Borenstein and Rose (1994) take up a similar question for the US airline industry. Since they observe more than two prices on a given airline/route, they use the Gini coefficient as a summary measure of price dispersion (rather than the price difference Δp for every product pair). They find that the expected price difference for two randomly selected passengers on a given airline/route is 36 per cent of the average ticket price. An increase in the number of competitors raises the extent of price dispersion by a large amount. Goldberg and Verboven (2001) measure margins based on the estimated own- and cross-price elasticities. They find that car manufacturers earn higher margins in their domestic markets than in their foreign markets, because markets are segmented according to country of origin and there is more competition in the foreign segments. Asplund et al. (2002) find that newspaper subscriptions in Sweden are more often sold at (often introductory) discounts in duopoly regions than in monopoly regions. They interpret this as evidence of poaching, that is, discrimination to attract new customers from rival firms.

The existence of indirect price discrimination is not obvious under competition, as shown in theoretical work. Nevertheless, empirical work has documented that competition may strengthen indirect price discrimination. Verboven (1999) finds a significant percentage price premium for optional engine power in the more competitive car segments, and percentage discounts in the less competitive segments (the latter being consistent with a monopoly discrimination). Busse and Rysman (2005) compare the prices of large and small ads in Yellow Pages directories. Their identification strategy relies on the assumption that cost differences between the two types of ads do not depend on the degree of competition. As such, they do not measure the extent of price discrimination per se, but instead ask how it varies with competition. They find that competition raises the discounts to large buyers: adding one competitor lowers the price of small ads by only six per cent, whereas it lowers the price of large ads by 12 per cent.

Economic Effects of Price Discrimination

Several empirical studies have also assessed the economic implications of price discrimination for profits, consumer welfare, tax revenues and economic efficiency. Leslie (2004) considers monopoly price discrimination. He finds that direct price discrimination for a Broadway theatre play, in the form of a currently observed 50 per cent discount at the discount booth known as the TKTS, raises profits five per cent above the profits under a uniform price strategy. However, he also finds that the current 50 per cent discount is too large to maximize profits, thereby generating too much substitution out of the full-price tickets. Lowering the discount to 30 per cent would raise profits seven per cent above the profits under a uniform price strategy. Leslie also estimates the aggregate consumer welfare effects from price discrimination, and finds them to be relatively small.

Under competition, the effects of direct price discrimination on profits are ambiguous even if the discriminatory prices are chosen optimally. The possibility to discriminate may lead to a situation of all-out competition, in which all discriminatory prices are lower than the uniform prices, thereby reducing profits. This occurs when the weak (elastic) market of one firm is the strong (inelastic) market of the other firm. Nevo and Wolfram (2002) find suggestive evidence of all-out competition, documenting that price discrimination (in the form of coupons) may lower the prices of all products, and may hence lower profits. Besanko et al. (2003) consider a situation of uniform pricing (for ketchup), and compute the new equilibrium under the assumption that firms would be able to discriminate between three (latent) customer segments. They find that all firms perceive the same customers as weak or strong. Price discrimination thus does not lead to all-out competition; quite the contrary, it increases profits. Brenkers and Verboven (2006) consider the reverse case in which car manufacturers currently discriminate between consumers from different countries, and would no longer be able to do this in the future (because of improved market integration). In their application, all-out competition appears more likely a priori, since domestic and foreign firms have the reverse strong and weak markets. Nevertheless, they find no evidence of all-out competition: an elimination of price discrimination would lower the prices of domestic firms, but raise the prices of foreign firms. Price discrimination correspondingly has relatively modest effects on industry profits and welfare (unless the high prices in the United Kingdom would be due to collusion).

The effects of indirect price discrimination under competition have also received attention recently. Miravete and Roller (2003) find that a single two-part tariff achieves 94 per cent of the potential profits and 63 per cent of potential welfare under a fully nonlinear tariff. McManus (2004) assesses the extent to which coffee shops distort their qualities (cup sizes) from the efficient levels, as a way to segment customers based on willingness to pay for quality. Consistent with economic theory, he finds that there are quality distortions, tending towards zero for the top qualities. Crawford and Shum (2007), using a somewhat different approach, also find evidence of quality degradation in the cable television industry.

See Also

- ▶ Price Discrimination (Theory)

Bibliography

Asplund, M., R. Erikkson, and N. Strand. 2002. *Price discrimination in oligopoly: Evidence from Swedish newspapers*. Discussion paper no. 3269. London: CEPR.

Benston, G. 1964. Commercial bank price discrimination against small loans: An empirical study. *Journal of Finance* 19: 631–643.

Besanko, D., J.-P. Dubé, and S. Gupta. 2003. Competitive price discrimination strategies in a vertical channel using aggregate retail data. *Management Science* 49: 1121–1138.

Borenstein, S. 1991. Selling costs and switching costs: Explaining retail gasoline margins. *RAND Journal of Economics* 22: 354–369.

Borenstein, S., and N. Rose. 1994. Competition and price dispersion in the U.S. airline industry. *Journal of Political Economy* 102: 653–683.

Busse, M., and M. Rysman. 2005. Competition and price discrimination in *Yellow Pages* advertising. *RAND Journal of Economics* 36: 378–390.

Brenkers, R., and F. Verboven. 2006. Liberalizing a distribution system: The European car market. *Journal of the European Economic Association* 4: 216–251.

Clerides, S. 2002. Book value: Intertemporal pricing and quality discrimination in the U.S. market for books. *International Journal of Industrial Organization* 20: 1358–1408.

Clerides, S. 2004. Price discrimination with differentiated products: Definition and identification. *Economic Inquiry* 42: 402–412.

Crawford, G., and M. Shum. 2007. Monopoly quality degradation and regulation in cable television. *Journal of Law and Economics* 50: 181–219.

Degryse, H., and S. Ongena. 2005. Distance, lending relationships and competition. *Journal of Finance* 60: 231–236.

Goldberg, P., and F. Verboven. 2001. The evolution of price dispersion in the European car market. *Review of Economic Studies* 68: 811–848.

Graddy, K. 1995. Testing for imperfect competition at the Fulton fish market. *RAND Journal of Economics* 26: 75–92.

Leslie, P. 2004. Price discrimination in Broadway theatre. *RAND Journal of Economics* 35: 520–541.

Lott, J., and R. Roberts. 1991. A guide to the pitfalls of identifying price discrimination. *Economic Inquiry* 29: 14–23.

McManus, B. 2004. *Nonlinear pricing in an oligopoly market: The case of specialty coffee*. Mimeo. Olin School of Business, Washington University.

Miravete, E., and L. Roller. 2003. *Competitive nonlinear pricing in duopoly equilibrium: The early U.S. cellular telephone industry*. Discussion paper no. 4069. London: CEPR.

Nevo, A., and C. Wolfram. 2002. Why do manufacturers issue coupons? An empirical analysis of breakfast cereals. *RAND Journal of Economics* 33: 319–339.

Scherer, F., and D. Ross. 1990. *Industrial market structure and economic performance*. Boston: Houghton Mifflin Company.

Shepard, A. 1991. Price discrimination and retail configuration. *Journal of Political Economy* 99: 30–53.

Verboven, F. 1999. Brand rivalry and market segmentation, with an application to the pricing of optional engine power on automobiles. *Journal of Industrial Economics* 47: 399–425.

Verboven, F. 2002. Quality-based price discrimination and tax incidence – The market for gasoline and diesel cars in Europe. *RAND Journal of Economics* 33: 275–297.

Price Discrimination (Theory)

Eugenio J. Miravete

Abstract

Price discrimination comprises a wide variety of practices aimed at extracting rents from a base of heterogeneous consumers. When consumer types are private information and only their distribution is known to the monopolist, finding the optimal nonlinear tariff involves solving a constrained variational problem that characterizes the optimal markup for each purchase level so that consumers of different types have no incentive to imitate the behaviour of others. Fully separating equilibrium is ensured when the distribution of types fulfills the increasing hazard rate property and individual demands can be unambiguously ranked. Outside this framework, optimal tariffs are difficult to characterize.

Keywords

Arbitrage; Exclusive agency; Incentive compatibility constraints; Individual pricing; Inverse elasticity rule; Market segmentation; Mechanism design; Mirrlees, J.; Multidimensional variational problem; Multiple products; Nonlinear pricing; Nonlinear taxation; Peak-Load pricing; Pigou, A.; Price discrimination; Ramsey pricing; Reservation price; Sequential screening

JEL Classifications

L10; C63; D43; D82

A monopolist price discriminates when he sells two identical units of a good at different prices, either to two different buyers or to the same customer. Two basic elements serve to classify the numerous methods whereby firms price identical units of the product differently: the amount of information available to the seller regarding how different the valuations of consumers are, and the seller's ability to avoid arbitrage. Avoiding arbitrage when firms sell personal services is easy and inexpensive, and thus price discrimination becomes a common practice in such industries. Conversely, in the absence of restrictions on the transferability of commodities, low-valuation customers could certainly benefit from reselling to higher-valuation customers, thus effectively impeding the seller from actually charging two different prices for the product.

Classification of Price Discrimination Practices

Pigou (1922) distinguished between first-, second-, and third-degree price discrimination depending on the amount of information regarding consumers' preferences that is available to the seller. In the case of first-degree price discrimination, the seller observes the actual valuation of each consumer and, provided that individual pricing is feasible, he could ask each consumer for her individual reservation price. Individual pricing is, however, rarely observed in reality, but such a pricing strategy has the theoretical appeal of leading to the efficient competitive outcome, although obviously with a quite different distribution of rents. This efficiency result vanishes when the seller knows only the distribution of consumers' valuations, as in second-degree price discrimination, or when he knows even less – just a signal about consumers' valuations – as in the third-degree price discrimination case.

Market segmentation, either geographical or personal, may serve as a way to avoid arbitrage. Price differentials across countries are likely to be larger than across neighbourhoods of a city as consumers move more freely in the latter case. Thus, the ability to price-discriminate will be partially determined by the importance of consumers' transaction costs in purchasing from different markets. Similarly, the cost of enforcing market segmentation may lead to different pricing schemes. Charging different individuals a different price for a service depending on their location, age, gender or race is far less expensive in terms of monitoring costs than tying prices to the income of each individual. In some circumstances, when third-degree price discrimination is used, location, age, gender, race or any other observable characteristics can be used in an economically efficient (although sometimes morally rotten) way to infer average individual valuations of products and increase profits by extracting a larger share of the consumer surplus of those individuals with higher valuations. Thus, in the third-degree price discrimination case, solving the price discrimination problem comes down to finding the optimal monopoly price in several independent markets. If there were numerous firms instead of a single seller, the well-known inverse elasticity rule should be modified to account for the existence of substitute goods.

More interesting is the case of second-degree price discrimination, when the seller needs to avoid the possibility of transferability of demand among consumers of different valuations. Since only the distribution of valuations is known, and not the valuation of individual consumers, finding the optimal pricing scheme requires one to solve a complex problem where the monopolist attempts to extract as much rent as possible from each consumer while at the same time ensuring that they do not imitate the behaviour of other consumers with lower valuations. In other words, price discrimination becomes a mechanism-design problem where a nonlinear tariff charging a different unit price for each unit sold maximizes the expected profits of the monopolist, while ensuring incentive compatibility.

Technical Issues of Single-Dimensional Price Discrimination

To solve this problem, consumers' preferences are assumed to be fully described by $U(q, \theta)$ where q represents the amount of good purchased by a consumer of type θ. This single-dimensional index captures the relevant difference in demand of diverse consumers, and leads to non-price-related shifts of individual demands. Type θ

remains private information for each consumer while the monopolist knows only its distribution $F(\theta)$ on $\Theta = \left[\underline{\theta}, \overline{\theta}\right]$. The variational problem that the monopolist faces consists of finding the optimal nonlinear tariff function $T(q)$ that maximizes his expected profits with respect to the distribution $F(\theta)$ provided that in their choices consumers are guided to maximize the net utility $U(q, \theta) - T(q)$. A fully separating equilibrium exists whenever individual demands can be ranked unambiguously, $Uq\theta(q, \theta) > 0$, and when the distribution of consumer types $F(\theta)$ fulfills the common increasing hazard rate property (these are sufficient, not necessary, conditions). In such a case, the optimal nonlinear tariff $T(q)$ is a concave function leading to quantity discounts that assigns different quantities and payments to consumers of different types. Maskin and Riley (1984) and Mussa and Rosen (1978) (in a framework of quality discrimination) first fully characterize the solution to this canonical version of the price discrimination problem. Contrary to the first-degree price discrimination case, now only the highest consumer type, θ, is efficiently priced – the efficiency at the top result – while all other consumers are charged a positive markup that induces them to self-select the optimal level of consumption according to the intensity of their preferences, θ. The magnitude of this markup depends on how difficult is to enforce the incentive compatibility condition, which is summarized by the hazard rate of the distribution $F(\theta)$. And the difficulty of separating different consumers depends on how these consumer types are distributed. Thus, the more numerous the consumers with a high valuation are, the larger is the average markup that low-valuation consumers should face in order to minimize the incentive of high-valuation types to purchase a small amount of the good. Intuitively, the more numerous high-valuation consumers are, the more likely some of them will be to attempt to behave as low-valuation consumers. To prevent it, a higher markup charged to low-valuation consumers is needed in order to reduce sufficiently the outside option of those more numerous high-valuation consumers. Consequently, if all consumers are alike, the distribution of consumer types, $F(\theta)$, becomes degenerate, and the optimal nonlinear tariff is a two-part tariff with a slope equal to the marginal cost of production and a fixed fee equal to the individual consumer surplus of a buyer.

Engineers (Dupuit 1849; Hadley 1885) rather than economists discovered long ago the advantages of charging different prices to different customers in order to cover the fixed costs of operating transportation services. The solution to the second-degree price discrimination problem described above attracted the attention of economists only after the contribution of Mirrlees (1971) in the area of nonlinear taxation. His approach to finding the optimal tax that maximized a social welfare function could easily be adapted to analyse the Ramsey pricing problem of regulated industries contemplated by Ramsey (1927) and Boiteux (1956). With the development of incomplete information games, the nonlinear pricing problem was rapidly reformulated as a direct revelation mechanism (Goldman et al. 1984; Guesnerie and Laffont 1984), thus helping to uncover the technical assumptions that ensured well-behaved solutions of the canonical single-product, single-parameters case.

Extensions of Monopoly Pricing

The solution to this canonical price-discrimination problem serves as a point of departure for many extensions that have attempted to incorporate either a more general theoretical approach or particular features of specific industries where nonlinear pricing is used to cover fixed costs or to fulfill distributional objectives set by regulators.

A first extension included the possibility that income effects were non-negligible and that consumers could be risk averse. Effectively, this means that the net utility of consumers is not additively separable in payments. Extensions in this direction includes the work of Mirrlees (1976), Roberts (1979), and Wilson (1993, ch. 7).

Another early extension addressed the rationing of stochastic individual demands in the presence of capacity constraints. The nonlinear tariff now attempts to distribute the cost of installed capacity among consumers according to their usage, as consumers with different loads contribute differently towards the cost of providing the service. But this peak-load pricing solution also provides the firm with incentives to reduce the size of consumers' loads in order to minimize the cost of distributing efficiently the existing capacity among all consumers. Oren, Smith and Wilson (1985) and Panzar and Sibley (1978) are the two basic references on capacity pricing.

More recently, the basic canonical model of price discrimination has been modified to contemplate the possibility of sequential screening, a process common in many industries where consumers first subscribe to one of the many optional tariffs available and later decide on their optimal level of consumption. The canonical model is modified to allow consumers to learn about their valuation of the product, thus distinguishing between *ex ante* and *ex post* types – the valuation of customers before and after contracting with the seller – as well as *ex ante* and *ex post* incentive compatibility constraints. Courty and Li (2000) consider the case where the *ex ante* type determines the distribution from where the *ex post* type will be drawn, while Miravete (1996) considers a framework in which the *ex post* type is the sum of the *ex ante* type and an independently distributed shock. Both approaches lead to ambiguous results that can be somewhat qualified depending on the stochastic dominance of the composition or convolution distribution, respectively, of the *ex post* relative to the *ex ante* valuation. Miravete (2005) further evaluates the welfare performance of nonlinear tariff options using data directly linked to *ex ante* and *ex post* types of consumers.

The most challenging extension of the canonical problem is multidimensional types. Wilson (1996) presents a concise description of the difficulties that arise when types are multidimensional or the monopolist sells several products. Type dimensions capture different features of consumer demands (intercept, curvature, or others) that are independent of prices but are relevant to capturing consumer heterogeneity. Multiple products introduce the possibility of accounting for complementarity and substitution effects and thus designing optimal discounts for bundles that include a variable proportion of each good. The difficulty arises because the multidimensional screening problem imposes a continuum of boundary conditions that translate into a large number (the number of type dimensions minus one) of additional partial differential equations that constrain the multidimensional variational problem. Explicit solutions do not exist beyond particular cases such as those studied by Armstrong (1996), Laffont, Maskin and Rochet (1987), or Wilson (1993, chs. 13, 14). A common result reported by Armstrong (1996) and Rochet and Choné (1998) is that low-valuation customers are always excluded, thus leading to bunching at the bottom.

Competitive Nonlinear Pricings

Besides the technical difficulties in solving multidimensional price-discrimination problems, numerical solutions show that tariffs may become non-monotone and that even the efficiency at the top result may not hold depending on the support of the distribution of types and the interaction among type dimensions given by the specification of the utility function. Perhaps because of these unsurmountable difficulties, the generalized one-dimensional nonlinear pricing framework of Rochet and Stole (2002) offers the most promising alternative for advancing in this area of research. Their approach consists of adding a second independent type dimension that enters additively into consumers' utility function. This little modification of the canonical problem disassociates the participation and consumption decisions. While in the canonical problem higher-valuation consumer types always participate and purchase more than low-valuation customers if lower types participate, now participation is driven by consumer-specific outside options. Now, relative to the canonical price-discrimination case, the monopolist loses some ability to extract consumer surplus as profit maximization requires him to balance informational rent extraction from high-valuation customers with the participation of low-valuation customers.

Characterizing the optimal tariff solution in this model with endogenous participation becomes more involved (although much more feasible) than the general multidimensional case, and it requires solving a two-point boundary problem instead of a simpler recursive first-order differential equation with a boundary condition given by the marginal consumer type that decides to participate in the market. Bunching may also occur at the bottom, but only at the bottom, and the tariff is well behaved, continuously approaching the fully efficient solution on the one hand and the solution to the canonical pricing problem on the other. Furthermore, the efficiency at the top result survives, and the efficiency at the bottom arises in cases where all consumer types are served.

The model of Rochet and Stole (2002) is also appealing because it offers the possibility of addressing competitive environments where firms' tariff are the best response to each other's offering and where the tariff offered by the competitor defines the outside option of consumers. This is a model of exclusive agency where consumers subscribe to only one of the firms competing in the industry. The most important result of this literature, also documented by Armstrong and Vickers (2001), is that, in industries with full market coverage and where all firms face the same marginal cost, the equilibrium tariff solution is a simple cost-plus tariff (Coasian two-part) leading to an efficient allocation of consumption among buyers.

See Also

- ▶ Mechanism Design
- ▶ Mechanism Design (New Developments)
- ▶ Price Discrimination (Empirical Studies)

Bibliography

Armstrong, M. 1996. Multiproduct nonlinear pricing. *Econometrica* 64: 51–75.

Armstrong, M., and J. Vickers. 2001. Competitive price discrimination. *RAND Journal of Economics* 32: 579–605.

Boiteux, M. 1956. Sur la gestion des monopolies publics astreint à l'équilibre budgétaire. *Econometrica* 24: 22–40.

Courty, P., and H. Li. 2000. Sequential screening. *Review of Economic Studies* 67: 697–717.

Dupuit, J. 1849. On tolls and transport charges. *International Economic Papers* 11(1962): 7–31.

Goldman, M., H. Leland, and D. Sibley. 1984. Optimal nonuniform pricing. *Review of Economic Studies* 51: 305–319.

Guesnerie, R., and J.-J. Laffont. 1984. A complete solution to a class of principal-agent problems with an application to the control of a self-managed firm. *Journal of Public Economics* 25: 329–369.

Hadley, A. 1885. *Railroad transportation: Its history and its laws*. New York: G.P. Putnam's Sons.

Laffont, J.-J., E. Maskin, and J.-C. Rochet. 1987. Optimal nonlinear pricing with two-dimensional characteristics. In *Information, incentives, and economic mechanisms*, ed. T. Groves, R. Radner, and S. Reiter. Minneapolis: University of Minnesota Press.

Maskin, E., and J. Riley. 1984. Monopoly with incomplete information. *RAND Journal of Economics* 15: 171–196.

Miravete, E. 1996. Screening consumers through alternative pricing mechanisms. *Journal of Regulatory Economics* 9: 111–132.

Miravete, E. 2005. The welfare performance of sequential pricing mechanisms. *International Economic Review* 46: 1321–1360.

Mirrlees, J. 1971. An exploration in the theory of optimum income taxation. *Review of Economic Studies* 38: 175–208.

Mirrlees, J. 1976. Optimal tax theory: A synthesis. *Journal of Public Economics* 6: 327–358.

Mussa, M., and S. Rosen. 1978. Monopoly and product quality. *Journal of Economic Theory* 18: 301–317.

Oren, S., S. Smith, and R. Wilson. 1985. Capacity pricing. *Econometrica* 53: 545–566.

Panzar, J., and D. Sibley. 1978. Public utility pricing under risk: The case of self-rationing. *American Economic Review* 68: 888–895.

Pigou, A. 1922. *The economics of welfare*. 4th ed. London: Macmillan.

Ramsey, F. 1927. A contribution to the theory of taxation. *Economic Journal* 37: 47–61.

Roberts, K. 1979. Welfare considerations of nonlinear pricing. *Economic Journal* 89: 66–83.

Rochet, J.-C., and P. Choné. 1998. Ironing, sweeping, and multidimensional screening. *Econometrica* 66: 783–826.

Rochet, J.-C., and L. Stole. 2002. Nonlinear pricing with random participation. *Review of Economic Studies* 69: 277–311.

Wilson, R. 1993. *Nonlinear pricing*. New York: Oxford University Press.

Wilson, R. 1996. Nonlinear pricing and mechanism design. In *Handbook of computational economics*, ed. H. Amman, D. Kendrick, and J. Rust, Vol. 1. Amsterdam: North-Holland.

Price Dispersion

Ed Hopkins

Abstract

Price dispersion occurs when different sellers offer different prices for the same good. Empirical studies have identified price dispersion as widespread and persistent. The most frequent explanation for this is that consumers do not have perfect information about prices. Only recently have economists succeeded in modelling price dispersion as an equilibrium phenomenon: that is, where consumers' decisions to acquire price information are a best response to the distribution of prices, and sellers' pricing decisions are a best response to consumers' search behaviour.

Keywords

Clearinghouse models; Price discrimination; Price dispersion; Sequential search

JEL Classifications

C7; D8

Price dispersion occurs when different sellers offer different prices for the same good in a given market. Thus, it differs from price discrimination under which a single seller offers different prices to different groups of buyers or in different geographical locations. A simple explanation for price dispersion is that it arises from imperfect information on the part of consumers, who do not all buy from the lowest price seller because some at least do not know who the lowest priced seller is. It is an important topic in the field of the economics of information in that there is considerable empirical evidence that price dispersion is widespread and significant. Yet it has proven surprisingly difficult for economists to derive satisfactory models that support price dispersion as an equilibrium phenomenon.

The rise of electronic commerce at the end of the 20th century gave new impetus to empirical studies of pricing behaviour. Baye et al. (2004) analyse detailed information on prices of 1000 items collected from a price comparison site. Price dispersion is found to be significant and persistent. Baye, Morgan and Scholten find an average coefficient of variation of about nine per cent for goods sold online. This is comparable with the results of Lach (2002) for conventional retailers who finds a lower coefficient for the price of refrigerators, but higher variation for grocery items such as coffee or flour.

Such empirical work on price dispersion is often disputed on the basis of two arguments, both of which claim that any apparent price dispersion is largely illusory. First, variance in prices might be explained by hidden heterogeneity in the good being offered for sale. For example, a retailer that charges high prices might survive not because of consumer ignorance of cheaper sellers, but because it offers superior service, something not captured by evidence on prices alone. A second line of scepticism is that dispersion in posted prices may not be inconsistent with uniformity in prices actually paid. Those who post high prices may not in fact sell anything. Certainly, one would expect low-priced sellers to sell more than those charging high prices, so that prices weighted by market share will be less dispersed than if all sellers are given equal weight.

The first criticism is addressed by Baylis and Perloff (2002) who find that, in fact, some online sellers persistently offer both high prices and poor service. The second is answered at least in part by Baye et al. (2004) who in their empirical study concentrate on the difference between the lowest and second lowest price, rather than the difference from lowest to highest or standard deviation, as their measure of dispersion. Furthermore, their data comes from a price comparison site where listings are costly for sellers. Why pay to list a price at which you think there will be no sales?

In any case, it is certainly possible to construct theoretical models in which prices are dispersed and yet high prices yield positive sales. Such theory is recent, however. In an influential survey, Rothschild (1973) identified serious difficulties with the then existing models of price dispersion. At that time, no one had produced a model where price dispersion was shown to be the result of equilibrium behaviour. The challenge was to show that charging a range of prices could be a rational response by sellers to the search behaviour of consumers, and vice versa.

It took some years for this challenge to be met. The difficulty in doing so is illustrated by the earlier work of Diamond (1971), who found that once one introduces imperfect information for consumers, a natural outcome is not price dispersion, but monopoly pricing by sellers. The essence of Diamond's result is the following. Suppose there are a large number of identical buyers who each want to buy one unit of a good from one of a large number of identical sellers, provided it costs no more than a maximum price p^*. The buyers know the distribution of prices but each only knows the price currently being charged by one seller. Each must then must decide whether to learn the price of one more seller at a fixed cost (imagine searching on foot, or by telephoning a succession of sellers). The optimal search policy in this situation of sequential search is to buy the first time one sees a price that is equal or below a reservation level r, which varies with the unit search cost s and distribution of prices $F(p)$. Now, if all consumers have the same unit search cost, then for a given distribution of prices, they will have the same reservation price r. The optimal price for all sellers must then be r. But if there is no dispersion in prices, it cannot be optimal to learn more than one price. Thus, the only equilibrium is where all sellers charge p^* and all buyers do not search, even when the unit cost of search is arbitrarily small. Ironically, this equilibrium satisfies Rothschild's criteria. Consumer behaviour is optimal since, when prices are identical, paying to learn additional prices is a waste of effort; pricing at the monopoly level is optimal since, when there is no search, there is no incentive for sellers to cut prices to increase sales.

Not surprisingly, therefore, many of the earliest successful equilibrium price dispersion models (Salop and Stiglitz 1977; Varian 1980) take a different route from Diamond and do not assume sequential search. Instead, they are what have been called by Baye et al. (2004) 'clearinghouse' models. By buying a newspaper or by visiting a price comparison website, a consumer can obtain information about the prices of a number of sellers all at once. The simplest clearinghouse assumption is that it is possible for consumers to become informed of all current prices. Suppose a proportion q of consumers remain uninformed and hence pick a seller at random. The other $1 - q$ consumers are informed and only purchase from the lowest priced seller. All consumers wish to buy one unit of the good if the price does not exceed a common maximum price p^*. Then, given n sellers and L consumers, if one seller charges a price strictly lower than all others, she sells to both informed and uninformed, a total of $qL/n + (1 - q)L$. The other sellers sell only to the uninformed and expect sales of qL/n. That is, demand is decreasing but discontinuous in price.

For simplicity, let us follow Varian (1980) and assume that sellers have constant marginal cost c. There is then no pure strategy equilibrium for sellers as long as there are both informed and uninformed consumers, that is if $q\ A\ (0,1)$. To see this, note that if all sellers charged the same price, it would generally be profitable for an individual to undercut this price in order to attract the informed buyers. However, because of the presence of uninformed consumers who are not price sensitive, charging the monopoly price p^* gives a guaranteed minimum profit of $(p^* - c)qL/n$, and so when the prices of other sellers are close to c, the most profitable price may be p^*. There is a symmetric mixed equilibrium in which all sellers randomize according to the same continuous distribution. This mixed equilibrium is a dispersed price equilibrium, because since sellers randomize over the prices they charge, realized prices will vary over sellers.

However, to have an equilibrium that fully satisfies Rothschild's challenge, it is necessary to make the consumer's decision to become informed endogenous. Varian (1980) assumed differing information costs, with high-cost consumers remaining uninformed, and low-cost consumers paying for information. However, Burdett and Judd (1983) showed that it is possible to close a clearinghouse model even with identical buyers. For example, given the symmetric mixed equilibrium described above, consumers who pay to become informed will buy from the lowest-priced seller whose expected price is equal to the expected value of the lowest of *n* independent draws from the equilibrium price distribution. In contrast, those who remain uninformed expect to pay the simple expectation of the distribution. If *q* is zero or 1, the equilibrium price distribution will collapse on *c* or *p** respectively. However, for interior values of *q*, the difference between the price paid by the informed and uninformed will be positive. Thus, it can be shown that for a value of *s* sufficiently low, there is at least one interior value of *q* such that the resulting equilibrium distribution of prices is sufficiently dispersed such that consumers are indifferent as to whether they pay or remain uninformed.

That is, there is at least one internally consistent dispersed price equilibrium. The proportion of informed consumers generates exactly the right amount of expected price dispersion such that consumers are indifferent between being informed and uninformed. This is an elegant but delicate construction. In contrast, the Diamond outcome (no consumers pay to be informed, all firms charge *p*)* is a simple pure equilibrium of this game that coexists with any dispersed price equilibria. Thus the Varian model and the similar models of Salop and Stiglitz (1977) and of Burdett and Judd (1983) have multiple equilibria (though the Bertrand outcome where all consumers pay to be informed and all firms charge marginal cost is not an equilibrium here, since consumers have no incentive to pay to be informed if all prices are the same).

A reasonable question is whether introducing heterogeneity, either under sequential search or in clearinghouse models, makes dispersed price equilibria more robust. However, consumer heterogeneity does not remove the Diamond paradox as an alternative equilibrium. Even if consumers have a range of search costs, if there is no price variation at all, then there is no incentive to search (unless one makes the implausible assumption that a mass of consumers have zero search costs). That is, if all sellers share the same monopoly price, then all charging that price can be an equilibrium if consumer search is costly. But if instead there is sufficient seller heterogeneity, an outcome where all sellers charge their monopoly price may not be an equilibrium. Suppose no consumer searches, each seller would then charge his or her monopoly price. However, suppose all consumers have the same continuous increasing demand function (in contrast to the unit demand assumed up to now), then a dispersion of costs amongst sellers would lead to heterogeneity in monopoly prices. This could be sufficiently diverse so that consumers would have an incentive to search. Thus, in the equilibrium of Reinganum (1979), low-cost sellers charge their monopoly price, but high-cost sellers must charge less than their monopoly price to make sales.

Finally, when one has heterogeneity of both buyers and sellers, there are two advantages. First, by the above argument, a Diamond-type outcome cannot be an equilibrium and so uniqueness of the dispersed price equilibrium is possible (Benabou 1993). Second, the dispersed price equilibrium can be pure and strictly monotonic: higher-cost firms charge higher prices. This follows because sufficient buyer heterogeneity can make demand to be continuous in prices, unlike the discontinuous demand in Varian's clearinghouse model. For example, if there is a continuum of buyers who search sequentially and have a continuous density of unit search costs, then there is the possibility of a continuous density of reservation prices. So, demand will increase smoothly as a seller lowers the price.

What are the major conclusions that can be drawn from these equilibrium models of price dispersion? The first is that both social and consumer welfare are typically decreasing with search costs. A reduction in search costs for some consumers can have a positive externality on other consumers, as increased search brings down prices for all. Other predictions can sometimes be counterintuitive. For example, an increase in the number of sellers actually raises the average price charged in the Varian model. However, this result does not hold for all price dispersion models. Further, Baye et al. (2004) find empirically that both average prices and the degree of price dispersion fall with an increase in seller numbers. Finally, we have seen that models with homogenous sellers give rise to mixed equilibria, while models with bilateral heterogeneity can generate pure equilibria. Randomization over prices would imply regular change in price order amongst sellers. That is, sometimes a given seller would have the highest price, sometime the lowest, and sometimes in the middle. A monotone pure equilibrium would give rise to a stable price ranking. Baylis and Perloff (2002) find that price ranking on online sales of electronic goods are very stable. In contrast, Lach (2002) finds that price ranking in data on prices charged by different Israeli supermarkets is highly variable.

One possibility is that the difference arises because Lach's data are for groceries that are purchased with greater frequency than the electronic goods in Baylis and Perloff's data set. But this highlights that the current theoretical literature on price dispersion has rarely addressed the related issues of repeat purchases, frequency of purchase and search patterns that depend on past experience, for example returning to sellers that have had low prices before. This would seem the area that is in most the need of further research.

See Also

- ▶ Oligopoly
- ▶ Price Discrimination (Theory)
- ▶ Search Theory

Bibliography

Baye, M.R., J. Morgan, and P.A. Scholten. 2004. Price dispersion in the small and in the large: Evidence from an internet price comparison site. *Journal of Industrial Economics* 52: 463–496.

Baylis, K., and J.M. Perloff. 2002. Price dispersion on the internet: Good firms and bad firms. *Review of Industrial Organization* 21: 305–324.

Benabou, R. 1993. Search market equilibrium, bilateral heterogeneity and repeat purchases. *Journal of Economic Theory* 60: 140–158.

Burdett, K., and K. Judd. 1983. Equilibrium price dispersion. *Econometrica* 51: 955–969.

Diamond, P. 1971. A model of price adjustment. *Journal of Economic Theory* 3: 156–168.

Lach, S. 2002. Existence and persistence of price dispersion: An empirical analysis. *Review of Economics and Statistics* 84: 433–444.

Reinganum, J.F. 1979. A simple model of equilibrium price dispersion. *Journal of Political Economy* 87: 851–858.

Rothschild, M. 1973. Models of market organization with imperfect information: A survey. *Journal of Political Economy* 81: 1283–1308.

Salop, S., and J. Stiglitz. 1977. Bargains and rip-offs: A model of monopolistically competitive price dispersion. *Review of Economic Studies* 44: 493–510.

Varian, H. 1980. A model of sales. *American Economic Review* 70: 651–659.

Price Level

P. Bridel

Until the end of the 19th century, it may be said that the quantity theory was everybody's theory of money and the price level. This does not mean that it was universally accepted: many writers submitted Hume's formulation to some very sharp criticisms. However, short of any viable alternative, all the leading economists adhered to one or another of the marginally different versions of the quantity theory.

The common feature of early-19th-century classical and late-19th-century neo-classical quantity theory is the well-known notion that an expansion or a contraction of the money supply – other things equal – would lead to an equiproportional change in the price level (or alternatively to an equiproportional change in the value of money). That) 'other things equal' is reflected in the assumption of a stable demand for money function, or, more specifically, a fixed level of output. The similarities between the Classical and Neo-classical approaches come however to an end here. Whereas in the latter approach the fixed (full employment) output assumption, and hence the causal relationship between money and prices, is the result of a theoretical analysis of the determination of output along marginalist lines, in the former it results from the adoption of Say's

Law. In other words, Classical quantity theory is based not on a theory of output but on the lack of such a theory comparable with its theory of value of distribution.

Accordingly, and despite attempts made by some of its leading proponents (like Thornton) to work their way toward a monetary analysis of the economic process as a whole in which price-level issues fall into secondary place, the Classical monetary theory, up until and including J.S. Mill, gave the pride of place to the so-called 'direct mechanism'. This) 'transmission mechanism' is older than economic theory itself. Much earlier than Hume's classical version, and well before economics was born as an independent subject, the idea that a change in the money supply would eventually cause prices to rise in the same proportion was part and parcel of most writings on money.

Even if Hume and Cantillon paid great attention to the manner in which a cash injection is disbursed and to the various lags involved in the process, and although they were well aware that an increase of money raises prices equiproportionately only if everyone's initial money holdings are increased equiproportionately, their attempts to prove it were thwarted by the very logic of the Classical framework. It is only with the Neoclassical effort to integrate money and value theories that the first serious attempts were made (mainly by Marshall and Wicksell) to escape from this Classical dichotomy and to prove the proportionality theorem by providing a proper stability analysis. However, and at least up until the early 1940s, most economists kept arguing that people spend more money because they receive more cash, not because the value of their real balances has increased beyond the amount determined by the Marshallian *k*. With his path-breaking analysis of the real-balance effect, Patinkin finally connected people's increased *flow* of expenditures with their feeling that their *stock* of money is too large for their needs. The sweeping endorsement of this theoretical argument by the economics profession allowed an apparently successful counter-attack against Keynes's claim that a fully competitive economy could well get trapped in (unemployment) disequilibria. Despite serious divergences among macroeconomists about the actual workings of the real-balance effect, it was widely held that, if prices and wages are flexible, a Walrasian equilibrium (with a positive value for money) would exist both in the short run and the long run. These investigations also confirmed that money is neutral; that is, excluding all distributional effects, in a neoclassical model coupled with unitelastic expectations, a once-and-for-all scalar change of all agents' initial cash holdings would change in the same proportion the equilibrium of money prices and nominal money balances at the end of the period, leaving unaltered relative prices and real variables. Price and wage rigidities are thus the only reasons that, in the short run, the excess demands for goods and money might not be homogeneous of degree zero and one respectively, with respect to nominal prices and initial balances.

The 'indirect mechanism' has a history that until the interwar period played second fiddle to that of the 'direct mechanism'. It is only with Marshall's, Wicksell's and, later on, Fisher's attempts to give an explicit rôle to the rate of interest in the transmission mechanism connecting money and prices that it rapidly took pride of place in the economist's monetary toolbox. In fact, the argument that monetary equilibrium (and hence the stability of the price-level in an economy) exists only when the money rate in the loan market equals the rate of return on capital (the traditional 'natural' rate) in the capital market is the basic framework within which some of the most famous discussions in the realm of monetary theory took and are still taking place. In all these analyses in terms of saving and investment, cumulative process, Gibson's paradox, forced saving, trade and credit cycles, etc, the price-level plays a crucial rôle as an indicator of the degree of tension within the system. Hence the wealth of introductory chapters on index numbers found in most textbooks and treatises of that period (the most famous being Book II in Keynes's *Treatise on Money* [1930], 1971).

With his cumulative process, Wicksell was indeed the driving force behind the impetus given toward the very end of the 19th century to this 'trailing rate' doctrine. Building on Tooke's 1844 insights, and in contradiction to Ricardo's pronouncements, Wicksell argued that, following a credit expansion, the market rate of interest and the price-level are positively correlated. As a matter of fact, the discrepancy (created by such a credit expansion between the market rate and the expected yield on investment) is a disequilibrium situation in which, period after period, net investment is positive and constantly increasing. Such a cumulative process need neither create inflation if voluntary savings is simultaneously generated via higher market rates (unless a 'pure credit' hypothesis is made) nor be explosive (thanks to the internal drain on banks' reserves). However, in order to preserve price stability, if the economy is operating at full employment and/or if there are signs of inflation, the bank rate would have to be raised in order to ensure that net investment does not exceed voluntary savings. Hence, a stable price level would not only be synonymous with equality between the real (or) 'normal') rate of return on investment and the market rate, but also with equality between the market and the bank rates. As Robertson put it later very succinctly:

It is on the difference between [Saving and Investment] and consequently between 'natural' and market rates that the movement of the price-level . . . depends' (1933, p. 411).

Within such a framework there began nearly half a century of intensive theorizing in terms of Wicksell's three criteria. The market rate is in equilibrium if it is equal to the rate of return of capital (or 'natural' rate), at which: (i) the demand for loans is equal to the supply of savings; and (ii) the price level is stable ([1896] 1936, pp. 192–9).

If the market rate trails behind the 'natural' rate, prices will begin to move up; if, furthermore, the bank rate diverges from the market rate, this creates an additional discrepancy between the market rate and the real rate of return on investment: the rate of inflation would of course gather up speed. In macroeconomic terms, the whole of this argument was ultimately incorporated in the loanable-funds theory of interest: the market rate of interest is determined by the demand for (investment demand and demand for cash balances) and the supply of loanable funds (voluntary savings and bank credit). If planned savings are equal to planned investment, net credit creation is equal to the demand for cash, the market rate, the bank rate and the 'natural' rate of interest are one and the same thing and, last but not least, the price level is stable.

Marshall in his stability analysis of the value of money (1923), Fisher in his famous equation (1911), Hawtrey with his purely monetary theory of the cycle (1913), Robertson with his) 'four crucial functions' (1928, pp. 105–7 and 182), most members of the Stockholm School (notably Myrdal 1929), Keynes in his famous 'fundamental equations' (1930) and Hayek (1932) with his forced saving analysis (to name only but a few of the most celebrated contributors to this debate) all tried, by putting a different emphasis on the various components of this indirect mechanism to offer a dynamic analysis of the price level. Having thus added money to a relativeprice system in which it has, by definition no part to play, these theorists tried in a certain sense to 'eliminate' it again by defining the monetary policy best suited to make money 'neutral' as concerns the operation of the economic system. By defining the prerequisites for money to be 'neutral', these authors were clearly implicitly (and sometimes explicitly) taking for granted the stability of the system. Rigidities, lags and inelasticities of all types, external shocks and technical progress, and of course monetary impulses could temporarily disrupt the dominant forces at work in an economy; but, ultimately, in the long run, the system would tend towards a full employment equilibrium along Walrasian lines. As Keynes wrote in his *Treatise:* 'Monetary theory, when all is said and done, is little more than a vast elaboration of the truth that "it all comes out in the wash"' (1930, II, p. 366).

Thus by the early Thirties and despite a great deal of activity in the field of monetary theory, the simple and straightforward) 'direct' and 'indirect' transmission mechanisms traditionally used to determine the price level were superseded because they proved, as Hayek argued, 'a positive hindrance to further progress' (1931, pp. 3–4). However, the rich harvest of new formulations and the stepping stones laid down by Hayek, a handful of Swedish economists and Hicks in the field of temporary equilibrium sustained no further work after the publication of Keynes's *General Theory of Employment Interest and Money.* With his *magnum opus,* Keynes simply changed the agenda of questions economists were to think about in the next thirty years. In particular, the central part played by the price level as the indicator *par excellence* in the course of the cycle was relegated together with the quantity theory to caricatural classroom teaching.

When Friedman resurrected the quantity theory as a theory of demand for money rather than as a theory of the price level (1950, p. 52), his intentions were originally to develop an alternative to the Keynesian liquidity preference argument. However, by asserting that the demand for money function was *empirically* stable and that it is *autonomously* determined, monetarist economists were able to relate again directly nominal income and price changes to changes in the stock of money. Friedman was thus explicitly in a position to consider his contribution as a theory of the aggregate price level the purpose of which is to provide the missing equation in a Walrasian system (1970, p. 223). The neo-classical synthesis having reached not too dissimilar conclusions, the Monetarist *vs* Keynesian controversy was ultimately seen by both sides as a debate on IS–LM elasticities, speed of adjustment and rigidities. In other words, and to quote Friedman, 'the fundamental differences between [these two streams] are empirical not theoretical' (1976, p. 315). All this suggests of course not only that there is an accepted theory of the economy but also that this theory is capable of yielding both monetarist and other conclusions. In other words, disagreements seem only to arise as far as the speed at which the economy converges to long-run equilibrium is concerned. Besides the fact that it is by no means the case that the IS–LM cross is a generally accepted theory of the economy (the Walrasian story monetarists see behind these two curves would certainly bar them from having income as one of their arguments), the assumptions one finds in the monetarist and New Classical Macroeconomic literature about the neutrality of money are not particularly plausible, let alone theoretically verifiable. In particular, they do not imply the uniqueness of such an equilibrium. Theorists like Hahn (1982) and Grandmont (1983) have shown that there are many, mostly a continuum of rational expectation equilibria over a finite horizon and there may also be many for an infinite horizon. Thus the belief that the long-run equilibrium of a competitive monetary economy is unique and stable and that a scalar change in the quantity of money holdings will generate the same scalar change in all nominal values remains today more than ever at the centre of a formidable theoretical debate. If the neo-classical monetary paradigm has survived, it is more because many economists *think* it yields important insights into the working of decentralized economies than for its theoretical solidity. Hence, and despite the empirical stability of the money demand function reported by many applied economists, and according to the maxim that what is witnessed if not explained is not understood, a proper *theory* of the price level remains yet to be written.

See Also

- ▶ Natural Rate and Market Rate of Interest
- ▶ Neutrality of money
- ▶ Quantity theory of money

Bibliography

Fisher, I. 1911. *The purchasing power of money.* New York: Macmillan. 1922.

Friedman, M. 1950. The quantity theory of money: A restatement. In *The optimum quantity of money*, ed. M. Friedman, 51–67. London: Macmillan. 1969.

Friedman, M. 1970. A theoretical framework for monetary analysis. *Journal of Political Economy* 78: 139–238.

Friedman, M. 1976. Comments on Tobin and Buiter. In *Monetarism*, ed. J. Stein, 310–17. Amsterdam: North-Holland.

Grandmont, J.M. 1983. *Money and value*. Cambridge: Cambridge University Press.

Hahn, F.H. 1982. *Money and inflation*. Oxford: Blackwell.

Hawtrey, R. 1913. *Good and bad trade*. London: Constable.

von Hayek, F. 1931. *Prices and production*. London: Routledge and Kegan Paul.
von Hayek, F. 1932. A note on the development of the doctrine of ‘forced saving’. *Quarterly Journal of Economics* 47: 123–33.
Keynes, J.M. 1930. *A treatise on money*, vol. 2. London: Macmillan. 1971.
Marshall, A. 1923. *Money, credit and commerce*. London: Macmillan.
Myrdal, G. 1929. *Monetary equilibrium*, 1939. London: Hodge.
Robertson, D. 1928. *Money*, 2nd ed. Cambridge: Nisbet and Cambridge University Press.
Robertson, D. 1933. Saving and hoarding. *Economic Journal* 43: 399–413.
Wicksell, K. 1896. *Interest and prices*. London: Macmillan. 1936.

Price Revolution

J. E. C. Munro

Abstract

The Price Revolution was a unique period of inflation in European economic history, enduring for 130 years, from the early 16th to the mid-17 century. It was fundamentally monetary in origins and character, having commenced with a fivefold increase in silver supplies from the central European mining boom and then sustained and expanded both by a financial revolution in negotiable credit instruments and then by the great influx of silver from the Spanish Americas. The extent of the inflation was, however, influenced by various real factors, especially demographic, which had their greatest impact on the income velocity of money.

Keywords

Bodin, J; Cambridge cash balances equation; Coinage debasement; Commodity money; Consumer price index; Deflation; Demography; Income velocity of money; Industrial revolution; Inflation; Malestroit, J; Money; Population growth; Price revolution; Spanish-American silver; Urbanization

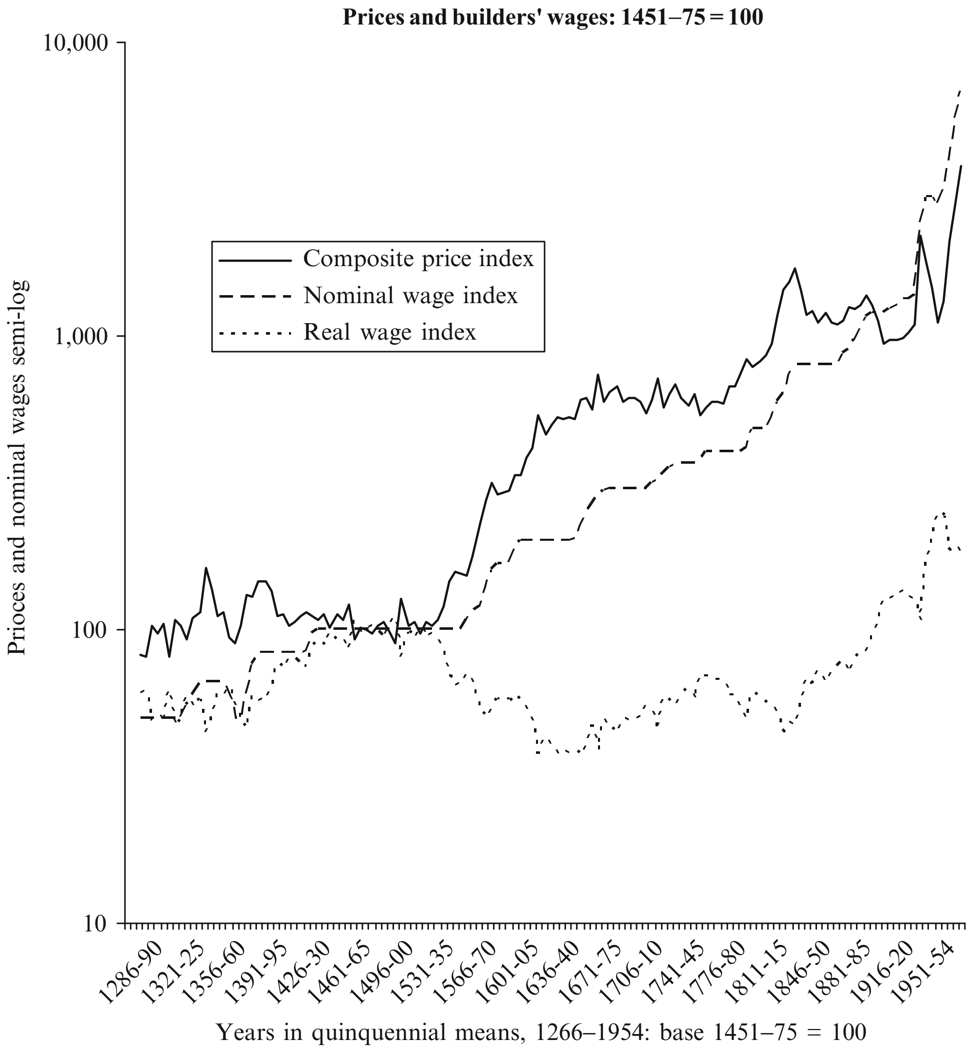
JEL Classifications

N3

The Price Revolution, dating from about 1515 to the 1650s, was a long period of persistent inflation in Europe that was unique for the pre-20th-century economy. The sustained rise in prices, or rather in the Consumer Price Index (CPI) is clearly visible in Fig. 1 for English prices from 1266 to 1954 (base 1451–75 = 100), and in Fig. 2 for prices in southern England, the southern Low Countries (Brabant), and Spain, from 1501 to 1650. With a common base of 1501–10 = 100 (CPI) for all three regions, we find that, over the next century and a half to 1646–50, the index number for Spanish prices rose to 343; for Brabantine prices, to 845; and for English prices, to 698.

Average annual rates of price increases of less than two per cent in the Price Revolution era may have been mild in comparison with 20th-century inflations: but all pre-20th century inflations were based on commodity moneys, not government issues of fiat money, as in the modern world. Before 1914, western Europe experienced, to be sure, other periods of long-term inflation, particularly, if only periodically, during the ‘long’ 13th century (1180–1315) and in the early Industrial Revolution era (1760–1815). But these produced price-level changes that were far smaller than those of the Price Revolution.

All long-term inflations are fundamentally monetary in nature, even though secondarily influenced by real factors. That may be best understood through the Cambridge cash balances equation, $M = k.P.y$, in which k indicates the quantity of cash balances (high-powered money M) held as a proportion of net national income (y). It is also the inverse of V, the income velocity of money, in the more familiar quantity-theory equation: $M. V = P.y$. Since the opportunity cost of holding cash is forgone interest income, changes in k should therefore depend partly on interest rates. Though an increase in M (money stocks) may prove inflationary, the equation indicates why the extent of such inflation is unpredictable. For such an increase in P can be offset by a rise in k (especially if an increased M reduces interest



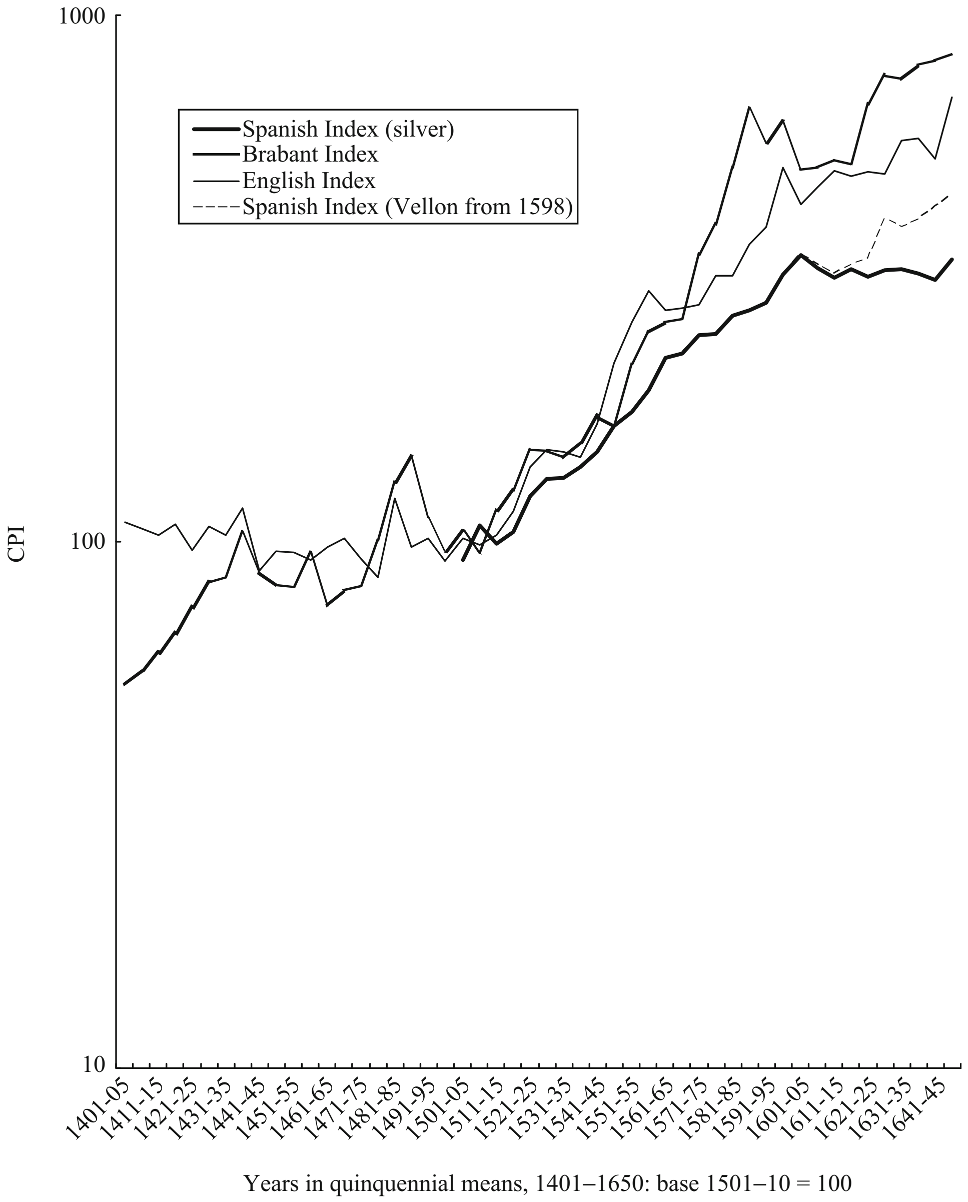
Price Revolution, Fig. 1 Consumer prices and wages for master building craftsmen in southern England, in quinquennial means: 1266–1954 (Phelps Brown and Hopkins indices). *Source:* Phelps Brown and Hopkins (1981, pp. 13–59)

rates), that is, a fall in *V,* or by an increase in y, stimulated by increased spending and falling interest rates.

Demographic Versus Monetary Explanations

Regrettably, amongst historians population growth has provided the most popular explanation for the Price Revolution. Contemporary explanations for the Price Revolution, especially in the debate between the French philosophers Jean Bodin and Jean Malestroit (1568), were instead purely monetary: that is, concerning the influx of Spanish-American silver during the 16th and 17th centuries. Modern opponents of this thesis have, however, rightly pointed out that virtually no American silver was imported before the 1530s, and only insignificant volumes were received before the 1560s, while inflation was clearly under way by 1515–20.

Yet to assume that consequently demographic factors provide the only possible alternative



Price Revolution, Fig. 2 Price indexes: England, Brabant, and Spain, 1401–1650. *Source:* England, as Fig. 1. Brabant, Van der Wee (1975, pp. 413–35). Spain, Hamilton (1934; pp. 262–403)

explanation is an absurd non-sequitur. There are two major problems with the demographic thesis. First, its most common form confuses microeconomic with macroeconomic changes. Although population growth, with fixed amounts of land and a static technology, should lead to a rise in the *relative* price of grains compared with those for manufactures, it can not explain a rise in the general price level. Second, the populations of both England and the Low Countries in the 1520s were at their late-medieval nadir, about half of what they had been around 1300 (when the English CPI was only 102); and thus any demographic recovery from such a low level could not possibly have provided the initial cause of an inflation that was under way in that very same decade.

The actual origins of the European Price Revolution lie instead in alternative monetary explanations, commencing with the central European silver-copper mining boom in the 1460s. This was an era of severe deflation (in silver-based prices) that had thereby augmented the purchasing power of silver and provided the key profit incentives for two crucial technological innovations: (*a*) in mechanical engineering: water-powered piston drainage pumps that permitted deeper mining, reaching richer ores; and (b) in chemical engineering, the *Saigerhütten* process using lead to smelt silver–copper ores, thus for the first time separating the two metals, which were present in vastly larger ore bodies than those of silver alone. The resulting silver–copper mining boom increased aggregate output of European mined silver about fivefold by the 1540s, producing far more silver than was imported from the Americas until the 1580s. By my own conservative estimates, central European silver production itself rose from an average annual of 12,873 kg in 1471–5 to 55,704 kg in 1536–40.

As late as 1556–60, only 27,145 kg of American silver were imported yearly into Seville; but in 1566–70 annual mean imports jumped to 83,274 kg, thanks to another technological innovation: the mercury amalgamation method, employed first at Potosi (Peru) and Zacatecas (Mexico). Thereafter, rising imports, reaching a maximum of 273,821 kg per year in 1591–5, but still amounting to an impressive 223,023 kg per year in 1621–5, continued to fuel the inflation. When the Price Revolution ended in 1656–60, silver imports had diminished to an annual mean of just 27,965 kg. Spanish-American mines were then experiencing severely diminishing returns, while far more metal was being retained for use in the Americas, and more and more silver was being exported across the Pacific, in trade with the Philippines and China.

There was one additional monetary factor to explain the European Price Revolution, namely, a veritable financial revolution in the Habsburg Netherlands, whose towns (from 1507) and then the Estates General (1539–43) established all the legal requirements for negotiability, including legalization of interest and discounting, to protect the rights of third parties in transferable bills, so that bills obligatory and bills of exchange could circulate from hand to hand in commercial and financial transactions as though they were paper money. This financial revolution also established full-fledged negotiability and thus far wider use of government debt instruments, internationally traded on the Antwerp *beurse* from 1531, as perpetual annuities known as *rentes* or *juros.* One measure of their vastly growing importance is the increased issue of Spanish *juros,* from 3.6 million ducats in 1516 to 80.4 million ducats in 1598, most of them held abroad. This financial revolution also increased the income velocity of high-powered money.

Demography and the Income Velocity of Money

Just the same, demographic factors are not irrelevant to our understanding of the dynamics of the Price Revolution, not when population growth became so much more dramatic from the 1540s to the 1640s. First, in various ways that have been elaborated by Harry Miskimin (1975), Jack Goldstone (1984) and Peter Lindert (1985), that population growth, combined with more urbanization, the development of more complex commercial and financial networks, and changes in the age pyramid (with more dependants), may have increased the income velocity of money. Furthermore, as Nicholas Mayhew (1995) has shown, the Keynesian predictions of a fall in income velocity with continued expansions in monetary stocks (and falls in interest rates) seems to hold true from the 13th to the 20th century, with one singular exception: the Price Revolution era.

The Role of Coinage Debasements

Finally, what explains the differences in the inflation rates revealed across the three countries in Fig. 2: why did Spanish prices rise less than

English, and English prices rise less than Brabantine? Coinage debasement (depreciation) seems to have played a role in these differences. Spain experienced no silver coinage debasement. England experienced one mild coinage debasement, in 1526, and one set of very severe debasements between 1542 and 1552 (though the silver coinage was only partially restored, in 1560–61); but none thereafter. Brabant, on the other hand, suffered a long series of coinage debasements, during the 16th and 17th centuries. Thus, the explanations for the European Price Revolution involve a complex set of monetary and real factors, though monetary factors predominated.

See Also

- ▶ Bodin, Jean (1530–1596)
- ▶ Commodity Money
- ▶ Cost-push Inflation
- ▶ Demand-pull Inflation
- ▶ Depreciation
- ▶ Economic Demography
- ▶ Fisher, Irving (1867–1947)
- ▶ Hyperinflation
- ▶ Inflation
- ▶ Inflation Measurement
- ▶ Keynes, John Maynard (1883–1946)
- ▶ Monetary Economics, History of
- ▶ Money
- ▶ Money Supply
- ▶ Population Dynamics

Bibliography

Goldstone, J. 1984. Urbanization and inflation: Lessons from the English Price Revolution of the sixteenth and seventeenth centuries. *American Journal of Sociology* 89: 1122–1160.

Gould, J. 1971. The price revolution reconsidered. In *The price revolution in sixteenth-century England*, ed. P. Ramsey. London: Methuen and Co.

Hamilton, E. 1934. *American treasure and the price revolution in Spain, 1501–1650*. Cambridge, MA: Harvard University Press.

Hamilton, E. 1971. American treasure and Andalusian prices, 1503–1660: A study in the Spanish Price Revolution. In *The price revolution in sixteenth-century England*, ed. P. Ramsey. London: Methuen and Co.

Lindert, P. 1985. English population, wages, and prices: 1541–1913. *Journal of Interdisciplinary History* 15: 609–634.

Mayhew, N. 1995. Population, money supply, and the velocity of circulation in England, 1300–1700. *Economic History Review,* 2nd series, 48: 238–257.

Miskimin, H. 1975. Population growth and the price revolution in England. *Journal of European Economic History* 4: 179–185.

Munro, J. 2003. The monetary origins of the ‘Price Revolution’: South German silver mining, merchant-banking, and Venetian commerce, 1470–1540. In *Global connections and monetary history, 1470–1800*, ed. D. Flynn, A. Giráldez, and R. von Glahn. Aldershot/Brookfield: Ashgate.

Outhwaite, R. 1982. *Inflation in Tudor and early Stuart England*, 2nd ed. London: Macmillan Press.

Phelps Brown, E., and S. Hopkins. 1981. *A perspective of wages and prices*. London/New York: Methuen.

Van der Wee, H. 1975. Prijzen en lonen als ontwikkelingsvariabelen: Een vergelijkend onderzoek tussen Engeland en de Zuidelijke Nederlanden, 1400–1700. In *Album offert à Charles Verlinden à l’occasion de ses trente ans de professoriat,* ed. Jan. Craeybeckx. Wetteren: Universum. Republished as ‘Prices and wages as development variables: A comparison between England and the Southern Netherlands, 1400–1700’, in H. Van der Wee, *The low countries in the early modern world*. Trans. L. Fackelman, Aldershot/Brookfield: Variorium- Ashgate Publishing, 1993, but without the statistical tables in the original Dutch publication.

Price, Langford Lovell Frederick Rice (1862–1950)

A. Petridis

Langford Price was born in London on 20 July 1862 and died in Brighton on 26 February 1950. He had a distinguished career as a student at Oxford, where he was elected to the first scholarship at Trinity College, graduating with firsts in Honours Moderations in 1882 and Literae Humaniores in 1885. Alfred Marshall lectured to Price at Oxford in 1885 and played an important part in his selection as first lecturer under the Toynbee Trust in 1885. Price held appointments as an extension lecturer at Oxford University until in 1888 he became a fellow and treasurer of Oriel

College, holding the fellowship until 1923. In 1907 he was appointed to the inaugural lectureship in economic history at Oxford, and in 1909 to the inaugural readership. He resigned from the readership on his retirement in 1921.

At the outset of his career Price had a conventional approach to economics, an approach greatly influenced by Marshall, who was both his teacher and mentor. However, after his appointment to the Toynbee Trust lectureship Price began to research into industrial relations problems in the shipbuilding industry on Tyneside, and his experience led him to gradually adopt a less orthodox approach to economics.

Price closely observed the operation of various wage systems (reinforced by interviews and correspondence with union officials and working men) and was persuaded that the competitive model of the labour market was unsatisfactory. Stimulated by the academic discussion and criticism of the published version of his Toynbee Trust lectures (*Industrial Peace,* 1887), Price published (1888) a unique analysis of short run bilateral monopoly which was one of the earliest attempts to grapple with aspects of the problem of bargaining. The 'model' postulated that in a bargaining situation there existed an upper and lower limit for the wage, these limits being established by competitive forces. Between these limits the actual wage is determined by bargaining power. The originality of Price's analysis at that time lay in his attempt to incorporate into the analysis of bargaining not only the direct costs but also the indirect costs of strikes which were to be set off against the costs of a particular wage settlement. Price's analysis was a precursor of modern bargaining theories and the theories of industrial disputes such as that developed by Sir John Hicks in the 1930s.

Price's questioning of the relevance of orthodox economic theory was reinforced by his teaching and research in economic history, and his growing identification with William Cunningham and William James Ashley in the dispute with Marshall over theory and facts in economic analysis. In 1903 Price made a final break from the Marshall-dominated laissez-faire school of economic theorizing. On 15 August 1903 *The Times* published a letter (the 'manifesto' as it came to be known) from a group of 14 leading economists who were opposed to the British government's protectionist proposals for tariff reform. Price was one of the few economists of non-professorial status asked to lend his prestige by signing the letter, but he undermined the 'manifesto' by declining to be a signatory and taking the unusual step of sending a copy of his letter of refusal to *The Times*. This letter was published immediately below the 'manifesto'. Price thus ensured maximum publicity for his dissenting views and his call for a Royal Commission.

After 1903 until his retirement in 1921 Price published only a few articles and participated only peripherally in professional societies, concentrating instead on his teaching in economics and economic history at Oxford. In those years Price filled the vacuum left by Edgeworth's lack of interest in the teaching of economics. Not only did Price teach economic history and history of economic theory courses each year, but he was active in the guidance and counselling of all students enrolled in the new diploma course in economics, which had been established in response to Price's representations to the Vice-Chancellor and the university community. In 1915 both Price and Edgeworth served on a committee which recommended the establishment of a new degree in economics, but the war prevented further action so that it was not until 1920 that a new honours school in Philosophy, Politics and Economics was established. Price publicly denounced the new honours school for not according a sufficient place to economics; despite his objections, the new degree structure for students who chose economics as their major subject was modelled on the diploma in economics Price had helped initiate and nurture through its early years.

Langford Price was never accorded full recognition for his original contributions to the analysis of bargaining under bilateral monopoly, for his strong and effective opposition to the use of the authority of economists to support the free trade position on tariff reform, and for his seminal role in the establishment of the teaching of economics at Oxford University. A complex of social, political and psychological factors would help to explain why he never received the accolades he richly deserved.

Selected Works

1887. *Industrial peace; its advantages, methods and difficulties*. London: Macmillan.

1888a. The relation between sliding scales and economic theory. *Section F, British Association for the Advancement of Science* 523–535.

1888b. 'West Barbary'; or notes on the system of work and wages in the Cornish mines. *Journal of the Royal Statistical Society* 51: 494–566.

1892. Notes on a recent economic treatise. *Economic Journal* 2: 442–462.

1896. *Economic science and practice*. London: Methuen & Co.

1902. *The present position of economic study in Oxford. (A letter to the Vice-Chancellor of the University.)* Oxford: published privately, Bodleian Library.

1904. Economic theory and fiscal policy. *Economic Journal* 14: 372–387.

1914. *Cooperation and copartnership*. London: Collins Press.

1937. *A short history of political economy in England*, 15th ed. London: Methuen & Co.

n.d. *Memoirs and notes on British economists 1881–1946*. Leeds: Brotherton Library.

Prices and Quantities

A. Brody

These are the most directly and readily observable attributes of commodities (goods and services produced for and exchanged on the market). Both price and quantity relate to a unit (piece, bushel, barrel, pound etc.), established usually by commercial practice as the customary unit of reckoning.

The intrinsically numerical character of prices and quantities renders accounts and statistics, the incessant measurement of the stream of commodities, feasible. This preoccupation is motivated by and yields motivation to business and economic interests. It also seems to be responsible for the profound drive to develop economic theories with the aid of mathematical tools, applied already successfully to the exigencies of natural sciences.

The units of measurement are manifold on the various markets and are also arbitrary to a certain extent. If the units undergo any changes, say, when measuring in grammes instead of ounces, then the numerical magnitude of both prices and quantities changes accordingly. Nevertheless this change in their numerical expression must not alter the total value (volume) of a given amount of commodities so measured: if the unit is doubled then the price of the new unit doubles likewise but the numerical expression of the quantity is halved.

This interdependence of prices and quantities prompted an historically early perception of their parallel, dual character. To this was soon added the appreciation of the mutual effect they exert on each other on the market. As Smith (1776) explained: if the quantity brought to market surpasses the effective demand, that is if an oversupply exists, this will depress prices. On the other hand, a high or excessively profitable price will induce a stepped up production of the commodity in question, possibly also a reduction in its effective demand. This skew-symmetric relationship, with quantities acting negatively on prices while prices influence quantities positively, has remained the popular wisdom of everyday economics up to the present day.

Later investigations and descriptions pointed to the existence of different mechanisms; be it the 'target farmer' in Third World countries who reacts to a rise in prices by reducing the quantity brought to market, or instances of administratively guided economic situations where the economic agents try to minimize their productive effort once prices are fixed. Still the basic form of interdependence on the market, as elucidated by Smith, remained valid in the majority of economic transactions and gained popular and scientific sanction and consensus.

Smith argued that there was a more or less perfect functioning of the 'invisible hand' of the market forces that promote equilibrium (equality of production and consumption, prices and costs) on almost all markets almost all of the time. Equilibrium therefore came to be seen as the normal state of affairs: the productive effort geared to match effective and solvent needs of the society. Random shocks, whether caused by changes in taste, technology or circumstances, were believed soon to be adjusted to. Hence the general prescription to economists (and politicians): not to interfere with this near perfect mechanism and not to tolerate obstacles, constraints, monopolies hampering the smooth operation of markets.

Here the economic profession split for the next two centuries. Economists less convinced about the fairness and impartiality, optimality and efficiency of markets and worried also about the historically emerging adverse tendencies, started critical investigations. They still accepted equilibrium as a theoretical tool of reasoning yet became increasingly aware of certain inadequacies observed on the market. With Ricardo (1817) and Marx (1867) the school of the labour theory of value came into being. This school maintained that prices and quantities are regulated in last instance by the respective amounts of live and congealed labour bestowed on the production of the commodities in question. They were interested mainly in long run tendencies in the economic circumstances of whole societies and used equilibrium reasoning to spell out these tendencies and also as a critical tool against existing imperfections. They were also responsible for developing more clearly the dual categories of value-in-use and value-inexchange: the extensive and intensive attributes of commodities. Marx particularly excelled in developing economic terminology in decidedly dual categories with analogous and parallel reasoning for price-type and quantity-type theorems as, for instance, the process of production and the process of realization, surplus product and surplus value, technical and organic composition of capital etc. This he considered as the main achievement of his approach.

> The best thing in my book is: 1. the emphasis on the dual character of labour, right in the first chapter, according to whether the labour is expressed in use value or exchange value (this is the basis of the whole understanding of facts).

The other school, less critical about the market and seeking rather the perfection of market mechanisms, has been interested more in short run responses of the economic system, looking for local and particular explanation of the actual behaviour found on the diverse markets. They maintained that prices and quantities are determined by the marginal adjustments needed to adapt to equilibrium; thus prices, in particular, depend on marginal costs and quantities will be determined by maximizing profits. Among others it has been mainly Pareto (1896) and Marshall (1920) who honed the economic arguments to the textbook precision of present day economics.

With Böhm-Bawerk (1896) the battle between the two schools became exacerbated and they spared no argument in refuting 'inimical' standpoints. This confrontation remained heated and mostly unjust on both sides, harbouring a sometimes implicit, sometimes explicit, political content roughly dividing the two camps into evolutionary and revolutionary protagonists.

Considering its strictly theoretical merits the feud, nevertheless, resembles the altercation in mechanics: Newton's followers starting from equilibrium considerations and in search of the *causa efficiens,* while d'Alembert's disciples fight for an optimizing approach and are looking for the *causa finalis,* the aim and purpose of motion. It took much time and pain to acknowledge finally the basic equivalence of the two seemingly inimical and antagonistic approaches.

A similar insight has been injected into economics by von Neumann (1937). The theoretical roots of his approach to and model of General Economic Equilibrium can be found partly in earlier unifying efforts in mathematical economics and partly in thermodynamic reasoning.

As a pioneer in mathematical economics Walras (1874–7) had already developed a model to determine the prices and quantities of a given economic system simultaneously. By establishing $2n$ equations in the $2n$ unknowns, n prices and

n quantities, he claimed the problem to be theoretically solved.

The idea was brilliant, the set-up ingenious, the proof incomplete. By counting equations it is not possible to prove existence and uniqueness of a mathematical solution. Even in the relatively simple case of linear equations where all the unknowns appear in their simplest form, multiplied only by some coefficients and then added up, the equations may be inconclusive. They may be contradictory, not permitting any solution at all. They may also be redundant and allow multiple solutions. And even if a solution exists and is unique we cannot exclude on *a priori* grounds some negative elements. Yet negative prices or negative quantities are usually meaningless in an economic context and cannot be accepted as genuine solutions.

These perplexing problems were eliminated finally by von Neumann in the following way.

Let $A = \{a_{ik}\}$ be the matrix of commodity inputs, $i = 1, 2, \ldots, m$ required to sustain one unit of the process $k = 1, 2, \ldots, n$ and similarly $B = \{b_{ik}\}$ the matrix of outputs yielded by the respective processes. Then, given p prices and x quantities (or 'intensities of production') pAx and pBx will express the total value of inputs (respectively, outputs). Thus $\lambda = \lambda\ (p, x) = pBx/pAx$ represents the rate of interest (as a relation of proceeds to advances in the process of realization, or the rate of possible growth as a relation of commodities produced to commodities consumed in the production process).

Analysing the gradients of this function leads to the following dual conclusion: If $\partial\lambda/\partial x = (pB - \lambda pA)/pAx$ is non-positive, that is if

$$pB \leq \lambda pA \tag{1}$$

then λ cannot be further increased by any variation of x and hence will be maximal. If inequality obtains in (1) for any k, then $x_k = 0$ because the process operates at a loss and should be discontinued.

If on the other hand, $\partial\lambda/\partial p = (Bx - \lambda Ax)/pAx$ is non-negative, that is if

$$Bx \geq \lambda Ax \tag{2}$$

then λ cannot be further diminished by any variation of p and hence will be minimal. If inequality obtains in (2) for any i, then $p_i = 0$ because the commodity is produced in a superfluous quantity and thus turns into a 'free' good.

Von Neumann now proved that the function λ (p, x) has a 'saddle point' for positive prices and quantities, where the maximal rate of growth equals the minimal rate of interest. Thus he succeeded in solving the economic problem of equilibrium by defining a so-called potential function and replacing equations by inequalities. Existence and positivity of prices and quantities in equilibrium still permit multiple equilibria, in a double sense.

Firstly, as can be seen, every multiple of the equilibrium price system yields the same equilibrium value and likewise every multiple of the equilibrium quantities is again a system in equilibrium. Thus only proportions and not absolute magnitudes are determined. Yet by choosing, as Walras did, one of the prices as 'numeraire' and expressing all the others as multiples of this 'numeraire' – and fixing one of the quantities as the reference unit – the system can be made wholly determinate.

Secondly, there are certain cases – they could be called) 'degenerate' – where true multiplicity of entirely different solutions may emerge. This problem can sometimes be remedied by a small perturbation of the initial data. Yet, it now appears that the possibility of multiple equilibria cannot be ruled out *ab ovo,* because they may appear in real economic systems just as well.

The theoretically decisive root of von Neumann's approach can be found in phenomenological thermodynamics, especially with Gibbs (1875), whose treatise 'On the Equilibrium of Heterogeneous Substances' synthesized classical thermodynamics and opened the way for physical chemistry. He applied first a 'max-min' criterion for equilibrium: maximizing entropy and minimizing energy, just as von Neumann maximized the growth rate and minimized the rate of interest, and he seems to have been the first to apply inequalities as well as equations in the description and analysis of equilibrium.

Von Neumann was fully aware of the analogy and stressed it when setting up his potential function Φ (X, Y) to be maximized by quantities X and minimized by prices Y:

> A direct interpretation of the function Φ (X, Y) would be highly desirable. Its rôle appears to be similar to that of thermodynamic potentials in phenomenological thermodynamics; it can be surmised that the similarity will persist in its full phenomenological generality (independently of our restrictive idealization).

Von Neumann's original notation followed the then accepted usage in physics: X for 'extensive magnitudes', that is quantities, and Y for 'intensive magnitudes', that is prices. The gradients of a potential function (the partial derivatives according to the variables) spell out the 'force field' in physics, and the vanishing of those gradients is the necessary requirement of equilibrium. In the von Neumann model, as in thermodynamics, theoretical considerations induce a complex) 'saddle point' problem: instead of simply maximizing the potential function the saddle point can be found only through minimizing by some and maximizing by other variables.

It is not pure coincidence that this thermodynamic approach proved to be so fertile in handling economic problems. New investigations in the axiomatic foundation of thermodynamics indicate (Giles 1964, p. 26) that 'any experimentally verifiable assertion of thermodynamics can be expressed in terms of states with the aid of the operation + and the relation → alone'.

Though the axioms related to the permitted → transformations may turn out slightly differently in economics – there is important work undertaken concerning variously formulated basic axioms, Debreu (1959) being a powerful and articulate example – it is evident that the mathematical structure underlying the two scientific disciplines is closely similar in each case.

The new approach, because of the unification of criteria of optimality with criteria of equilibrium, did much to bridge the gap between the two opposing schools of economic thought. Both found their basic ideas tolerably well reflected in the set-up of the von Neumann model and hence a new round of revision and even partial reconciliation could be started.

One should stress: it has been surely the 'restrictive idealization' that facilitated the general acceptance of the new approach. The model only encompasses linear processes with a linear combination of inputs, resulting in a likewise linear combination of commodities. It represents, furthermore, only the production of freely reproducible commodities, that is: it does not contain any external constraints on the scale of production. Such a model keeps data and computational requirements relatively modest and is also easy to grasp.

With matrix notation now universally accepted this convenient shorthand made the model mathematically transparent. The very simple statement of dual equilibrium: $\lambda pA = pB$ and $\lambda Ax = Bx$ could not possibly be simplified further.

We now have an almost complete mathematical theory of so-called 'matrix pencils', that is matrices of the form $A + \lambda B$. It is interesting to note that Weierstrass (1867) reported on his investigations concerning this form in the same decade in which most of the ingredients, indispensable for our topic to take its present shape, were published. Marx, Walras, Gibbs and Weierstrass made known their results in the same decade not only independently but without having the slightest notion about each other.

With the advent of computers, also pioneered by von Neumann, matrices with several thousands of rows and columns became manageable and this permitted and motivated an everbroadening use and proliferation of a family of models having their theoretical and mathematical source in the von Neumann model.

Some very important and justly famous models were developed in the next decades. Being all equivalent in a mathematical sense to the Neumann model, as it has been demonstrated in most instances by the respective authors themselves, they can and should be considered as mathematical variants of the latter: input-output analysis, as proposed by Leontief (1941), linear programming, as investigated by Dantzig (1947) and Kantorovich (1940), the neo-Ricardian model set up by Sraffa (1960) and finally two-person game theory, an earlier product of von Neumann (1928), reaching broader scholarly circles only with the von Neumann and Morgenstern (1944) volume. (The last contains a further generalization to *n*-person games.)

In spite of the mathematical equivalence those models have been developed mostly independently and have roots in widely different economic considerations. Sraffa's approach, a careful and consistent restatement of Ricardo's value theory, proved to be particularly important. The underlying idea, if possible, is even more simple here. In a self-replacing system where, in the absence of growth, $\lambda = 1$ with no joint products, hence $B = 1$, the prices can be determined unequivocally by the postulate: the inputs required to reproduce the respective commodities have to be defrayed from the proceeds of selling the same commodities. Hence the proportions of prices and quantities are determined by the dual system of equations

$$pA = p \text{ and } Ax = x \tag{3}$$

Still in the more realistic cases, when extended reproduction and joint products have to be admitted, the description and solution is more rigorously and easily furnished by embedding the Sraffa system in a general von Neumann model.

Considering also the neo-Marxian restatement of labour theory as furnished by Brody (1970) and Morishima (1973), exploiting the Leontief model, where

$$p(A + \lambda B) = p \text{ and} (A + \lambda B)x = x \tag{4}$$

and B interpreted as a stock-input matrix, a certain consensus seems to be reached:

According to the neoclassical exposition of Hahn (1982), all the schools would compute the same numerical magnitudes for prices and quantities for an economic system in equilibrium. They would accept the same system of equations, though they would interpret those equations differently. Deeper and yet unreconciled differences emerge only when abandoning the critical point of equilibrium.

With painfully won reconciliation in sight a new theoretical attack on equilibrium reasoning takes shape. Kornai (1971), collecting all the critical observations and deeply influenced by the inadequacies of economic systems which endeavour to replace the market by equilibrium computations declared: the equilibrium school 'has become a brake on the development of economic thought'.

Paradigms – and equilibrium thinking is one such, with a domain much broader than economics alone – are seldom damaged by criticism. They may be done away with only by new and more powerful paradigms. Hence they rather thrive on objections – and all the internal problems already emerged with Smith who implicitly or explicitly maintained that equilibrium (i) exists, is (ii) optimal, is (iii) pursued and is also (iv) achieved.

Existence has been proved yet under 'restrictive idealization' in linear models but by a shrewd mind, knowing that it is permitted to approximate most functions, however complicated, linearly by taking their derivatives in the neighbourhood of the point analysed. (This may be achieved by taking a series expansion and neglecting terms of higher order.) The isomorphism of matrices and operators has been also well known to the pioneer of operator theory. So it is no wonder that all the models introduced are wide open to further generalization. Here non-linear programming, with Kuhn and Tucker (1956) and Martos (1975) and non-linear input–output models with Morishima (1964) have to be mentioned, also the success in generalizing the Neumann model by Medvegyev (1984) and applying operator calculus with Thijs ten Raa (1983). An increasing unification with linear and non-linear systems theory and with modern non-equilibrium thermodynamics can be safely predicted.

Optimality has also ethical, social, psychological and political connotations because one has to propose an entity (growth rate, utility, satisfaction, equity etc.) to be optimized. In this respect our subject belongs to the domain of welfare economics. Mathematically, the question is fairly simple: equilibrium and optimality can be made to correspond because solving equations is equivalent with minimizing the errors of the solution. That is: the solutions of $Ax = b$ and $Ax = r$ with $\Sigma\,(r - b)^2 \rightarrow$ minimal are the same if they both exist.

Ethical, political, and other convictions will of course always influence scholars in choosing and developing their topics but, luckily, they do not play any role in proving or refuting theorems and corollaries.

Stability, the question whether equilibrium can or cannot be achieved, if pursued, and maintained, once achieved, is the most interesting question in the forefront of present research. The stability analysis of economic systems, performed by methods borrowed again from physics and also thermodynamics: analysis of the eigenvalues of the response matrix, negative definiteness, discussion of the second partial derivatives, the le Chatelier–Braun principle etc. indicate that both market and planning systems are usually stable, yet seldom asymptotically stable, and if asymptotically stable the speed of convergence is usually very slow.

Stability means that a given deviation from equilibrium will not grow without bound: if the deviation is initially small it will not become infinite. This secures the feasibility of the system, its ability to function; yet a system may be stable and perform very poorly. Even asymptotic stability, that is achieving the decline and vanishing of discrepancies, is an unsatisfactory criterion in economic matters because by the time the equilibrium point is reached or approximated it may be already displaced by changes of the system itself.

In reality economic systems move not in slowly changing equilibrium states but along socalled transients, a succession of non-equilibrium positions. Thus we are still far from an acceptable theory of economic motion. The models introduced spell out requirements of equilibrium but not the actual forces bringing, or not bringing the system to equilibrium. Still, certain inroads have been made by models of cycles, for example, Kalecki (1935), Goodwin (1967) and Brody (1985).

But perhaps more important than analysis seems to be the task of synthesis. Acknowledging that neither plan nor market can avoid economic fluctuations, the quest for controlling prices and quantities in a smoother and more efficient way is understandable. Questions of optimal control in linear and nonlinear systems emerge and once approximately solved the search will go unavoidably deeper: how to control the position of equilibrium itself, how to become master of structure and technology. To shape interdependence itself in a conscientious manner, to influence the outcome of technological and structural change is the next item on the agenda of mathematical economics.

See Also

- ▶ Command Economy
- ▶ Linear Models

Bibliography

Böhm-Bawerk, E. 1896. Zum Abschluss des Marxschen Systems. In *Festgaben für Karl Knies,* Berlin. Trans. *Karl Marx and the Close of his System* by Eugen Böhm-Bawerk, ed. P. Sweezy, Reprints of Economic Classics, New York: A.M. Kelley, 1966.

Brody, A. 1970. *Proportions, prices and planning*. Amsterdam: North-Holland.

Brody, A. 1985. *Slowdown: Global economic maladies*. Beverly Hills: Sage.

Dantzig, G. 1947. Maximization of a linear function of variables subject to linear inequalities. In *Activity analysis of production and allocation*, ed. T.C. Koopmans. New York/London: Wiley/Chapman, 1951.

Debreu, G. 1959. *Theory of value: An axiomatic analysis of economic equilibrium*, Cowles foundation monograph, vol. 17. New York: Wiley.

Gibbs, J.W. 1875. On the equilibrium of heterogeneous substances. *Transactions of the Connecticut academy* III, October 1875–May 1876 and May 1877–July 1878. (See also *The scientific papers of J. Willard Gibbs*, New York: Dover, 1961.)

Giles, R. 1964. *Mathematical foundations of thermodynamics*. Oxford: Pergamon Press.

Goodwin, R. 1967. A growth cycle. In *Socialism, capitalism and economic growth*, ed. C.H. Feinstein. Cambridge: Cambridge University Press.

Hahn, F. 1982. The neo-Ricardians. *Cambridge Journal of Economics* 6(4): 353–374.

Kalecki, M. 1935. A macrodynamic theory of business cycles. *Econometrica* 3(1): 327–344.

Kantorovich, L.V. 1940. Ob odnom effektivnon metode reshenia nekotorih klassov extremalnih problem (trans: On an efficient method to solve some classes of external problems). *Dokladi Akademii Nauk SSSR* 28.

Kornai, J. 1971. *Anti-equilibrium*. Amsterdam: North-Holland.

Kuhn, H.W., and A.W. Tucker (eds.). 1956. *Linear inequalities and related systems*, Annals of mathematics studies, vol. 38. Princeton: Princeton University Press.

Leontief, W. 1941. *The structure of American economy 1919–1929*. New York: Oxford University Press.

Marshall, A. 1920. *Principles of economics*, 8th ed. London: Macmillan, 1964.

Martos, B. 1975. *Nonlinear programming. Theory and methods*. Amsterdam: North-Holland.

Marx, K. 1867. *Capital*, vol. I. Moscow: Foreign Languages Publishing House, 1961.

Medvegyev, P. 1984. A general existence theorem for von Neumann economic growth models. *Econometrica* 52(4): 963–974.

Morishima, M. 1964. *Equilibrium, stability and growth: A multisectoral analysis*. Oxford: Clarendon Press.

Morishima, M. 1973. *Marx's cconomics. A dual theory of value and growth*. Cambridge: Cambridge University Press.

von Neumann, J. 1928. Zur Theorie der Gesellschaftsspiele (trans: On the theory of games of strategy). In *Collected Works.* Oxford: Pergamon Press, 1963.

von Neumann, J. 1937. A model of general economic equilibrium. In *Collected works*. Oxford: Pergamon Press, 1963.

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

Pareto, V. 1896. Cours d 'économie politique. In *Oeuvres complètes*. Geneva: Libraire Droz, 1964–7.

Ricardo, D. 1817. Principles of political economy and taxation. In *Works and correspondence of David Ricardo*, vol. 1, ed. P. Sraffa. Cambridge: Cambridge University Press, 1951.

Smith, A. 1776. *An inquiry into the nature and causes of the wealth of nations,* ed. E. Cannan. London: Methuen, 1961.

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

Thijs ten Raa. 1983. Dynamic input–output analysis with distributed activities. *IFAC/IFORS Conference Reprints,* Washington, DC.

Walras, L. 1874–7. *Elements of pure economics or the theory of social wealth*. London: Allen & Unwin, 1954.

Weierstrass, K. 1867. Zur Theorie der bilinearen und quadratischen Formen (trans: On the theory of bilinear and quadratic forms). Berlin: *Monatshefte der Akademie der Wissenschaften,* 310–338.

Prices of Production

G. de Vivo

This expression is used by Marx, mainly in Volume III of *Capital,* to indicate the exchange values of commodities, when he fully takes into account the inconsistency between the uniformity of the rate of profits and an exchange in proportion to labour embodied. He accordingly tables the famous 'problem' of the 'transformation of the Values of Commodities into Prices of Production' – i.e., into prices which would include profits at a uniform rate on *the whole* capital advanced ($c + v$), and which would therefore differ from relative embodied labours, when commodities differ in the 'organic composition of capital' (c/v). The expression 'prices of production' (which Marx regards as synonymous with 'cost-prices', and with Smith's 'natural prices') starts to be commonly used by Marx only some time after he had actually formulated the 'transformation problem', and provided his solution. In *Theories of Surplus Value* (1862–3), for instance, he would normally still employ its synonym 'cost-prices'. But 'prices of production' is used in the 1893 draft plans for Volume III of *Capital* (printed in Marx 1862–3, I, pp. 414–16).

Although the classical economists generally used the expression 'natural prices', 'prices of production' had some currency in the mid-1810s, when it was used by Torrens, in the first edition of his *Essay on the External Corn Trade* (1815, p. 229; at this time, it was also used by Malthus and Ricardo in their correspondence, but not in their published writings).

It is noticeable that Torrens not only preceded Marx in employing the expression 'prices of production', but also formulated something very similar to his 'transformation'. Torrens criticized Ricardo's labour theory of value on the ground that 'the rate of profit in the several occupations of industry always tends to an equality', and therefore 'as equal capitals generally put unequal quantities of labour in motion, . . . the products of equal quantities of labour will be of unequal value' (Torrens 1818, pp. 57–8). He accordingly states that the price of each commodity would be determined by adding profits at a uniform rate to the value (labour embodied) of the total capital employed in its production. This is basically what Marx does in his 'transformation'. The main difference is that Torrens does not determine the rate of profits as a ratio between the labour values of the profits and of the capital [$s/(c + v)$], and he is in general unable to determine it (for more details on Torrens's theory, see de Vivo 1986; see also Robbins 1958, p. 60 ff.).

Marx knew Torrens's 1815 *External Corn Trade*, which is quoted in Volume I of *Capital*. In Volume III of *Theories of Surplus Value* he discusses at length Torrens's conceptions on value, to some extent acknowledging Torrens's anticipation of his points (1862–63, III, p. 72).

See Also

- ▶ Centre of Gravitation
- ▶ Cost of Production

Bibliography

De Vivo, G. 1986. Torrens on value and distribution. *Contributions to Political Economy* 5.

Marx, K. 1862–3. *Theories of surplus Value*, vols I–II-I. London: Lawrence & Wishart, 1969–72.

Robbins, L. 1958. *Robert Torrens and the evolution of classical economics*. London: Macmillan.

Torrens, R. 1815. *An essay on the external corn trade; Containing an inquiry into the general principles of that important branch of traffic; an examination of the exceptions to which these principles are liable; and a comparative statement of the effects which restrictions on importation and free intercourse are calculated to produce upon subsistence, agriculture, commerce, and revenue*. London: Hatchard.

Torrens, R. 1818. Strictures on Mr Ricardo's doctrine respecting exchangeable value. *Edinburgh Magazine*, October. As reprinted in R. Torrens, *The economists refuted and other early economic writings*, ed. P. Groenewegen. University of Sydney: Reprints of Economic Classics, 1984.

Pricing on the Internet

Michael R. Baye and John Morgan

Abstract

While many conjectured that the information-rich and frictionless nature of online markets would result in marginal cost pricing, this has proved not to be the case. Price dispersion online is ubiquitous. The main reason is that price discovery occurs through platforms that have an incentive to ensure that prices are dispersed so that information is valuable. We survey models of platform pricing and trace the impact of their decisions downstream to e-retailers. Finally, we highlight the connection between empirical findings and theory predictions for e-retail pricing.

Keywords

E-retail; Internet; Network effects; Platform; Pricing; Price dispersion; Two-sided market

JEL Classification

D4; D8; M3; L13

Overview

Initial studies of pricing on the internet focused on e-retail pricing, where many conjectured that the frictionless nature of online markets would result in marginal cost pricing and the 'law of one price'. More recent attention has centred on the pricing decisions of platforms – websites such as Google, Amazon, eBay and Facebook – that are basins of attraction for consumers as well as firms. The prices charged by platforms and e-retailers are intertwined: a platform's choices feed into downstream advertising and pricing decisions, and vice versa. Of course, both sets of prices affect consumer browsing and buying decisions.

This article recognises these interconnections and begins upstream, at the platform level. We discuss the evolution of the literature on platform pricing, including access fees, transaction fees, menu prices and auctions. We then examine how market structure influences platform pricing. Next, we move downstream to discuss the pricing and advertising decisions of e-retailers, and conclude by examining how market structure shapes these decisions.

Platform Pricing

Baye and Morgan (2001) provide the first model of optimal pricing by a platform that serves consumers and firms in a two-sided market for information. In their model, $n > 1$ geographically separated towns are each serviced by a local firm. While transaction costs preclude consumers domiciled in one town from physically visiting stores in other towns, a third player – an independent platform – operates a virtual marketplace that can tear down the geographic barriers separating consumers and firms. Firms advertise at the platform to gain access to consumers in distant towns; consumers visit the virtual marketplace, gain access to the list of advertised prices, and benefit if they find a price better than that charged by their local firm. The platform recognises that information is a valuable resource, and charges access fees to consumers and firms using its site. The access fee on the buyer side of the market represents a consumer's cost of subscribing to (or accessing) the platform's website, while on the seller side it is an advertising or listing fee.

The two main findings are: (1) the platform's optimal access fees result in e-retail prices that are dispersed even though firms are identical, and (2) the platform finds it optimal to charge consumers low (or even free) access fees, instead profiting from advertisers. Finding (1) stems from the value proposition of the virtual market: absent price dispersion, consumers find the platform's information to be worthless and, absent consumers, firms find the platform itself to be worthless. The platform therefore endogenously injects frictions into the market (by charging positive access fees on the seller side of the market) to maximise its profits. Finding (2) stems from an externality between the two sides of the market: by charging consumers low access fees, the platform attracts consumers, creating a virtuous circle inducing firms to advertise. Baye and Morgan show that the platform captures more value by maximising consumer participation than from extracting rents from both sides of the market.

The term 'two-sided market', which never appears in Baye and Morgan, became prominent due to Rochet and Tirole (2003). In contrast to Baye and Morgan, who emphasise within-side competition among sellers using the platform, Rochet and Tirole study situations where such competition is effectively absent. Additionally, while Baye and Morgan focus on the access fees charged to participants on different sides of the market, Rochet and Tirole study transactions fees.

In the Rochet and Tirole model, a buyer and seller meet to determine whether to conduct their transaction through the platform or not. While they couch the model in terms of credit card use, it also applies to online dating and other match-based services. A consumer will use the platform if her value (v^b) exceeds the price (p^b) charged by the platform for the transaction. Let $D_b(p^b)$ denote this probability (referred to as quasi-demand). Analogous conditions hold for sellers with a transactions value of v^s and a transactions price p^s. Assuming v^i are statistically independent, and there is a constant marginal cost of processing each transaction, the platform's per-transaction profit is

$$\pi = \left(p^b + P^s - c\right) D_b\left(p^b\right) D_s\left(p^s\right)$$

Assuming that quasi-demands are log-concave, optimal pricing depends, in a simple way, on relative elasticities. The transactions price in each side of the market satisfies a variant of the standard monopoly markup formula, but accounts for the connection between the two sides. Specifically, the optimal transaction fee satisfies

$$p^i = \frac{1}{1 - \eta^i}\left(c - p^j\right)$$

where η^i denotes the quasi-demand elasticity for a particular side i of the market.

This implies that, other things equal, an increase in the elasticity of quasi-demand on either side of the market reduces the total transactions price.

The transactions prices paid by parties on different sides of the market are determined by the ratio of their elasticities:

$$\frac{p^b}{p^s} = \frac{\eta^b}{\eta^s}$$

Thus, the predicted pricing behavior has the feature that the side with the *more* elastic demand pays the *higher* transactions fee. This prediction is the opposite of that for conventional markets, and illustrates how optimal pricing rules in two-sided markets differ from standard results.

This pricing structure depends critically on the log-concavity assumption, which rules out commonly used empirical demand specifications such as the constant elasticity formulation arising when values are Pareto-distributed. For this case, Bolt and Tiemann (2008) show that the platform charges *lower* prices to the more price-sensitive side of the market. More precisely, a profit-maximising platform sets a sufficiently low transaction fee to induce full participation on the more elastic side of the market, and captures surplus by charging higher fees on the less elastic side. This structure is identical to Baye and Morgan's finding in information markets, where the platform is restricted to *access* fees.

The broad lesson is that optimal platform pricing is sensitive to the structure of demand on each side of the market, the set of pricing instruments available to the platform, and the nature of externalities within and between sides of the two-sided market. There is no simple or universal solution to the problem of optimal platform pricing – details matter.

Recognising this, subsequent research generalised these early models along two key dimensions: expanding the set of price instruments available to the platform and the nature of network externalities between the two sides of the market. Armstrong (2006) is notable in both respects. His linear utility structure allows for variability in the value of each additional platform user to the other side of the market. He also allows the platform to offer a combination of access and transaction fees. As with Rochet and Tirole, within-side competitive effects are absent, primarily for tractability. Armstrong studies competition between horizontally differentiated platforms.

Two key insights emerge from Armstrong's analysis. First, despite the presence of network effects, platforms can coexist in equilibrium provided they are sufficiently differentiated. Second, these externalities actually sharpen price competition. Compared to the situation where such network effects are absent, platforms offer lower prices both as a defensive response to aggressive pricing by rivals as well as offensively to attract market share that makes their network more valuable. In equilibrium, platform pricing depends on the degree of market power and externalities on both sides of the market.

Baye et al. (2011) also enrich the set of pricing instruments available to platforms. In contrast to Armstrong, their setting features within-side competition among sellers. Their main finding is that, despite having the possibility of using two-part tariffs (a combination of fixed access fees and variable transactions fees), a monopoly platform optimally prices solely through transaction fees. This result rationalises an important trend in platform pricing during the decade of the 2000s. At the start of the decade, platforms such as price comparison sites typically based their fees on impressions (eyeballs) rather than actions of users (clicks). By the end of the decade, the so-called CPC (cost per click) model of advertising was dominant. Under CPC, advertisers only pay when a transaction (a click) occurs. While this shift was widely seen as a concession to advertisers in the face of uncertain returns to online advertising, their result implies that platforms, in fact, benefit from this pricing practice.

Weyl (2010) further expands the set of pricing instruments available to a monopoly platform to include price menus that depend on the level of participation on each side of the market. While he allows for general demand and externalities in the market, he excludes within-side competitive effects. In this setting, Weyl demonstrates the optimality of *insulating tariffs* – contingent pricing where prices on one side of the market depend on the level of adoption on the other side. Such pricing structures are common in advertising at traditional media outlets. For instance, standard television advertising contracts contain provisions that adjust the rates paid by advertisers based on the number of viewers of the show on which the advertisement appears.

A separate strand of the literature abstracts from network externalities entirely and focuses instead on mechanisms to deal with the apparently insoluble pricing problems confronting platforms. For instance, to optimally price an advertisement triggered by a search query, the platform must account for the query itself, information about the user, the time and location in which the query takes place and so on. Rather than using a top-down method for pricing, Google and other search platforms have turned to auctions, essentially letting participants on the advertiser side of the market solve the pricing problem for them. These so-called 'slot auctions' have their own unique structure: advertisers place a single bid, indicating their maximum willingness to pay per click. Bids are then sorted from the highest to lowest with the highest bid getting the top slot (i.e. the highest position on the page of search results), and succeeding bids are allocated the next lowest slot until all slots are taken up. In practice, a combination of factors determines an advertiser's position. Roughly speaking, the platform estimates the clicks generated by a given ad, and multiplies this by the bid to determine expected revenues, which are ordered by bidder. (The exact process by which this determination takes place is a trade secret of each platform.)

Most slot auctions use the generalised second price (GSP) rule to determine payments. A simple form of this rule has the highest bidder pay the second-highest bid, the second-highest bidder pay the third-highest bid, and so on. This pricing rule seems similar to the familiar second-price auction mechanism, and thus would appear to induce advertisers to bid their true (private) values, as in Vickrey (1961). This is not the case, however; Edelman et al. (2007) and Varian (2007) show that, generically, truthful bidding is not an equilibrium outcome. Nonetheless, there exists an equilibrium to a GSP auction that produces the same expected revenues as a Vickrey auction.

Another important difference between slot auctions and standard auctions concerns revenue equivalence. The predecessor to the GSP auction, a form introduced by the firm Overture, had each bidder simply pay its own bid. Based on the revenue equivalence theorem, one might be tempted to conclude that this auction form produces the same expected revenues as the GSP auction. Edelman et al. show that this is not the case either; in general, the GSP auction outperforms the Overture auction.

Market Structure and Platform Pricing

We now turn to the impact of market structure on platform pricing. The network effects typically present on both sides of the market can easily lead to a situation of natural monopoly and, indeed, in many real-world platform markets, there is a single dominant player. Facebook is the dominant social networking site, Google dominates search, eBay dominates online auctions, and so on. Other markets, however, are more fragmented; for instance, no single dominant platform has emerged in online dating. This section explores how industry fundamentals influence market power, and hence pricing.

A key dimension along which the examples above differ concerns the degree of horizontal differentiation. For dating platforms, horizontal differentiation is paramount. Indeed, one of the more popular platforms, Jdate, expects its users to be Jewish, which obviously limits the size of its potential market. By contrast, horizontal differentiation seems less important for the choice of operating systems. Competition in Rochet and Tirole (2003, 2006), Armstrong (2006), Armstrong and Wright (2007), and others is generated through sufficient horizontal differentiation among platforms. But what if such differentiation is absent?

Caillaud and Jullien (2003) provide an answer to this question in an important early paper studying competition between vertically differentiated platforms. In their model, platforms compete in access and transactions fees and differ in the efficiency with which they match the two sides of the market. Competitive effects on a given side of the market are absent in their setting. Their main finding is that monopoly structures emerge although the threat of entry limits the market power of the monopolist. This is consistent with the widely held view that platform coexistence is unstable when platforms are undifferentiated or vertically differentiated.

Ellison and Fudenberg (2003) and Ellison et al. (2004) challenge this view.

They point out that, while a platform with the larger market share generates scale effects (users of an online auction site benefit from greater breadth of offerings, for instance), within-side competition on a platform creates a countervailing *market impact effect* (additional sellers on an online auction platform lead to lower seller payoffs). They show that the market impact effect can be large enough to offset the scale effect and allow platforms of very different size to coexist.

A simple example captures their intuition. Suppose that an online auction market consists of three sellers and six buyers choosing between two identical platforms. Platform A attracts one-third of the buyers and sellers in this market while platform B attracts the remaining two-thirds. By virtue of its size, buyers and sellers on platform B enjoy higher surplus than those on platform A. This begs the question: Why don't the individuals on platform A simply switch to platform B? The key is the market impact effect. A buyer switching from A to B increases the competition on platform B and, consequently, raises the price for the item since there are now 5 buyers rather than 4 competing for the same 2 items. With a large enough price increase, this can more than offset the scale advantage and cause buyers to remain at the smaller platform. A similar effect is present for sellers at the smaller platform.

The lesson from this literature is that the presence of within-side competitive effects fundamentally changes conclusions about market structure, even when platforms are undifferentiated.

Brown and Morgan (2009) empirically examine the implications of the Ellison et al. models. They conduct field experiments by selling rare coins on two competing online auction sites in the US: eBay and Yahoo. They find no evidence of compensating market impact effects, instead concluding that this market was in the slow-motion process of tipping to eBay. (Subsequent to their experiments, Yahoo closed its US online auction site, leaving eBay as the single dominant player.) Using laboratory experiments Hossain et al. (2011) also investigate the dynamics of platform competition, varying the degree of horizontal and vertical differentiation, as well as market impact effects. Regardless of the magnitude of the market impact effect, they find that platforms coexist only when there is sufficient horizontal differentiation; otherwise, the market tends to tip to the more efficient platform. The phenomenon of tipping to quality appears in many empirical studies as well. For instance, Tellis et al. (2009) investigate market shares across competing technology platforms. They find that the higher-quality platform (as reflected by review sites) tends to dominate its market.

E-Retail Pricing

We now examine the implications of platform pricing on downstream retailers' pricing and advertising decisions. We highlight two key features of the landscape: price dispersion and the trend toward posted prices rather than auctions.

The primary focus of the early e-retail literature concerned price dispersion. While price dispersion was widely observed in offline markets, the main explanation for it, dating back to the seminal paper of Stigler (1961), was search friction, including 'shoe-leather' costs. The internet dramatically reduced search friction – physical visits to stores were no longer necessary and extensive product information was readily available online. It stood to reason that price dispersion should vanish in an internet world, a view capably summarised by *The Economist* (20 November 1999, p. 112), which argued:

> The explosive growth of the Internet promises a new age of perfectly competitive markets. With perfect information about prices and products at

> their fingertips, consumers can quickly and easily find the best deals. In this brave new world, retailers' profit margins will be competed away, as they are all forced to price at cost.

This 'brave new world' turned out to be a fiction. Brynjolfsson and Smith (2000), in the earliest comprehensive study of online pricing, document considerable price dispersion – the range in online prices for an identical product often exceeded 100% for the books and CDs they surveyed. They conclude that '…while there is lower friction in many dimensions of Internet competition, branding, awareness, and trust remain important sources of heterogeneity among Internet retailers' (p. 563).

The price dispersion documented above was for listed prices, but consumers care about *landed* prices, including shipping, handling and taxes. Since these additional costs constituted a large portion of the landed price of books and CDs (still mainly sold in physical rather than in digital form at the time of their study), consumers would economise by bundling their purchases. This led some to argue that the law of one price, while violated on a product-by-product basis, might still hold once one properly accounted for bundles of goods.

Baye et al. (2004) address this concern by focusing on higher priced consumer electronics products such as computer monitors, cameras, printers and PDAs (personal digital assistants). Shipping costs for these items represent a small fraction of the price (which averaged about $500 in their sample). Consequently, bundling is less important to purchase decisions. They too found evidence of considerable and ubiquitous price dispersion. Using even the most conservative measure, the gap between the lowest and second-lowest prices on offer, they found average dispersion levels of 5%, a far cry from the world envisaged by *The Economist.* Moreover, they found systematic differences in price dispersion depending on the number of firms listing prices at the comparison site (platform) they studied. The gap between the two lowest prices systematically shrinks with the number of competitors, but the range in prices increases in the number of competitors. A key point raised in the study is that the effect of competition depends crucially on the measure of price dispersion being used.

The first theoretical model that formally rationalised online price dispersion is that of Baye and Morgan (2001). As discussed above, firms in this model can advertise their prices on the platform to attract geographically distant shoppers. Firms must pay the platform a fee to advertise, but will only succeed in attracting these shoppers when they offer the lowest advertised price. Firms face a tradeoff between charging low prices and paying fees to attract distant shoppers, or charging the monopoly price, eschewing the platform, and catering to local customers. Resolving this tradeoff entails mixed, or randomised, pricing and advertising strategies. These strategies make both price and geographic reach unpredictable, thereby preventing rivals from systematically offering better deals on the platform.

The essence of the model is captured in a simple environment where S shoppers use the platform to buy at the lowest price and L loyal consumers per firm do not.

All consumers have unit demand up to a reservation price r for an identical product sold by n firms with constant marginal cost m. When the platform charges firms an access fee of ϕ, each advertises on the platform with probability

$$\alpha = 1 - \left(\frac{n\phi}{(n-1)(r-m)S}\right)^{\frac{1}{n-1}}$$

A firm that does not advertise on the platform charges a price of r, while a firm that does advertise sets its price $p \in [p_0, r]$ based on the distribution function

$$F(p) = \frac{1}{\alpha}\left(1 - \left(\frac{\frac{n}{n-1}\phi + (r-p)L}{(p-m)S}\right)^{\frac{1}{n-1}}\right)$$

The lower bound of the distribution of prices is

$$p_0 = m + \frac{\frac{n}{n-1}\phi L(r-m)}{L+S}$$

Notice that, despite the fact that firms compete purely on price and are identical in every way, prices are always above marginal cost. More importantly, this pricing formula highlights the link between upstream platform pricing (ϕ) and downstream e-retail pricing: the higher is the listing fee charged by the platform, the higher are the resulting e-retail prices (in the sense of first-order stochastic dominance). Thus, the strategic injection of "frictions" by the platform itself in the form of a positive listing fee, $\phi > 0$, impacts pricing, with higher fees softening of price competition and raising firm profits.

Several other models explaining price dispersion are nested in this specification, including Shilony (1977), Varian (1980), Rosenthal (1980), Narasimhan (1988) and Iyer et al. (2005); see Baye et al. (2004) for details. The key forces highlighted in these models are differences in consumer information and loyalty.

Specifically, Varian (1980) argues that informational differences among consumers are critical drivers of price dispersion. In his model, some consumers are perfectly informed about available prices while others are uninformed and face high search costs. One can think of these uninformed consumers as those on the wrong side of the 'digital divide', i.e. those lacking access to the internet or unaware of the search tools available online.

Rosenthal (1980) argues that brand loyalty drives price dispersion. In his model, some consumers are loyal to a particular firm and view its services as superior to those offered by all other firms. The remaining consumers are shoppers, who view all sellers as identical and purchase from the firm offering the lowest price.

Iyer et al. (2005) extend the above models to include targeted advertising *à la* Butters (1977). In their model, firms advertise not simply to inform consumers about prices, but also to alert consumers to their very existence, and this again produces price dispersion via mixed strategies.

A parallel literature explored the unique features of auctions as a selling mechanism. This literature is less relevant to the current e-retail landscape as fewer goods are sold online via auction. Online auctions fundamentally transformed a number of markets, particularly for collectibles, once dominated by small stores. Even for these items, auctions are now less used – over 70% of eBay's revenues now come from sales via posted prices. Largely this is because the types of items sold online now more closely resemble the mix sold offline. In addition to digital media, such as books, music and games, traditional offline goods such as apparel and white goods are now offered and sold online. The need for price finding via an auction is largely absent in markets for goods where a 'street price' may be readily ascertained. Of course, auction models are still useful for modeling some posted-price markets; see Spulber (1995).

We conclude by noting that, while many e-retailers responded to the plethora of information available online by cutting prices, today there is a renewed emphasis on non-price competition. This includes providing customers with high levels of service, fast shipping, one-stop shopping, seamless returns and so on. In terms of value creation, the greatest impact of the internet has arguably been the ease with which consumers can locate products. Brynjolfsson et al. (2003) proved prescient in this regard; they document that even in the very early days of the online book market, most of the value creation stemmed from consumers' abilities to find books in the 'long tail'.

Market Structure and e-Retail Pricing

Most people view e-retail as being extremely competitive, so presumably individual sellers face highly elastic demands. But just how elastic are the demands faced by e-retailers? Chevalier and Goolsbee (2003) use a clever identification strategy to provide an answer for online booksellers. By directly intervening in the book market and observing how their purchases affect a book's sales ranking, they deduce the demand elasticities faced by the two largest booksellers. They estimate Amazon's elasticity to be about 0.6 while Barnes & Noble's is about 4. In contrast, Ellison and Ellison (2009) report an elasticity in excess of 20 for a firm selling aftermarket RAM (computer memory) – much closer to perfect competition.

Baye et al. (2009) suggest that differing market structures for books and aftermarket RAM memory can account for these differing elasticity estimates. A few large players dominate the market for books, while the memory market is highly fragmented, with hundreds of small firms competing. Standard pricing models indicate that a firm's elasticity of demand varies with the number of its competitors; thus, one might expect more elastic firm demand for RAM than for books. They examine this hypothesis using clicks data on PDAs sold through Kelkoo, a price comparison site in the UK. When only a single seller lists a price, the estimated elasticity is around 2, but it rises to 6 when 15 sellers list prices. Thus, market structure appears to be an important determinant of an e-retail firm's demand.

One of the key implications of the models described earlier, which feature shopper/loyal or informed/uninformed consumer segments, is that a firm experiences a discontinuous jump in its demand when offering the lowest price on the platform. A firm that cuts its price from slightly above to slightly below the low-price competitor captures the price sensitive consumer segment. Baye et al. (2009) estimate that a firm experiences a 60% jump in demand in these situations, which is consistent with a situation where about 13% of Kelkoo's users are price sensitive 'shoppers'.

Another important feature of these models is that some consumer segments may not benefit from competition. For instance, in the Varian model, informed consumers expect to pay the *minimum* listed price, which falls with the number of competing firms owing to an order-statistic effect. In contrast, uninformed consumers pay the average listed price, which *rises* as more firms compete. The intuition for this perverse result is that, by reducing the chance that a firm will attract informed consumers, competition discourages firms from cutting prices. Thus, increased competition disparately affects the prices paid by consumers in different segments: there are both winners and losers.

The ease of price comparison and ready access to product information has turned some e-retail markets, such as aftermarket RAM, into landscapes resembling perfect competition. Firms have countered by developing a number of strategies to inject informational frictions back into the search process, thereby softening price competition. One such strategy is obfuscation (or, more bluntly, bait and switch). Here, an e-retailer offers a low-price/low-quality item at the platform and then attempts to upsell the consumer who clicks through to the e-retailer's own site. Ellison and Ellison (2009) note the effectiveness of such strategies in the computer memory market.

Another strategy is to break a price into various parts, such as a base price and a shipping charge, and make one of these parts (shipping, usually) difficult to ascertain. This strategy, termed price 'shrouding' by Gabaix and Laibson (2006), is increasingly prevalent among e-retailers. Gabaix and Laibson provided a theoretical rationale for such strategies, which Hossain and Morgan (2006), as well as Brown et al. (2010), explore empirically. They conducted field experiments in online auction markets for music CDs, video games and iPods, varying both the opening bid and the shipping charge. They find that consumers do not fully account for the shrouded aspects of prices, and that merchants can (and do) profit from these cognitive errors. In a large-scale follow-up study using eBay data, Levin (2011) confirms these results.

We conclude by noting that the rise of vertical integration makes it increasingly difficult to distinguish e-retailers from platforms. Leading this trend is Amazon, which began life as an upstart online bookseller, but has evolved into a retail platform featuring its own offerings, offerings of affiliated merchants, and even outside offerings determined through bidding, much like a search engine. Apple, which initially focused on hardware, is now the largest digital music platform. Its App Store represents an enormous software platform, and its aggressive move into e-books triggered a fierce fight with Amazon (and competition authorities) over its pricing practices. Google, meanwhile, has evolved from being purely a search engine to providing mobile operating systems, browsers and mobile hardware (through its acquisition of Motorola). How these changes affect online pricing represents an important open question. Chen (2008) provides an interesting initial look at how vertical integration by an 'information gatekeeper' into the product market it serves affects upstream and downstream pricing decisions.

See Also

- ▶ Auctions (experiments)
- ▶ Auctions (theory)
- ▶ Bertrand Competition
- ▶ Electronic Commerce
- ▶ Hotelling, Harold (1895–1973)
- ▶ Internet and the Offline World
- ▶ Location Theory
- ▶ Online Platforms, Economics of
- ▶ Price Dispersion
- ▶ Product Differentiation
- ▶ Search Theory (New Perspectives)
- ▶ Search Theory
- ▶ Two-Sided Markets
- ▶ Vickrey, William Spencer (1914–1996)

Bibliography

Armstrong, M. 2006. Competition in two-sided markets. *RAND Journal of Economics* 37(3): 668–691.

Armstrong, M., and J. Wright. 2007. Two-sided markets, competitive bottlenecks and exclusive contracts. *Economic Theory* 32(2): 353–380.

Baye, M.R., and J. Morgan. 2001. Information gatekeepers on the internet and the competitiveness of homogeneous product markets. *American Economic Review* 91(3): 454–474.

Baye, M.R., J. Morgan, and P. Scholten. 2004. Price dispersion in the small and large: evidence from an internet price comparison site. *Journal of Industrial Economics* 52(4): 463–496.

Baye, M.R., J. Rupert, J. Gatti, P. Kattuman, and J. Morgan. 2009. Clicks, discontinuities, and firm demand online. *Journal of Economics & Management Strategy* 18(4): 935–975.

Baye, M.R., X. Gao, and J. Morgan. 2011. On the optimality of clickthrough fees in online markets. *Economic Journal* 121(556): 340–367.

Bolt, W., and A.F. Tieman. 2008. Heavily skewed pricing in two-sided markets. *International Journal of Industrial Organization* 26(5): 1250–1255.

Brown, J., and J. Morgan. 2009. How much is a dollar worth? Tipping versus equilibrium coexistence on competing online auction sites. *Journal of Political Economy* 117(4): 668–700.

Brown, J., T. Hossain, and J. Morgan. 2010. Shrouded attributes and information suppression: Evidence from field experiments. *Quarterly Journal of Economics* 125(2): 859–876.

Brynjolfsson, E., and M.D. Smith. 2000. Frictionless commerce? A comparison of internet and conventional retailers. *Management Science* 46(4): 563–585.

Brynjolfsson, E., M.D. Smith, and Y. Hu. 2003. Consumer surplus in the digital economy: Estimating the value of increased product variety at online booksellers. *Management Science* 49(11): 1580–1596.

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

Caillaud, B., and B. Jullien. 2003. Chicken and egg: Competition among intermediation service providers. *RAND Journal of Economics* 34(2): 309–328.

Chen, J. 2008. Backward integrated information gatekeepers and independent divisions in the product market.B.E. *Journal of Theoretical Economics* 8(1), Article 7.

Chevalier, J., and A. Goolsbee. 2003. Measuring prices and price competition online: Amazon.com and Barnesand-Noble.com. *Quantitative Marketing and Economics* 1: 203–222.

Edelman, B., M. Ostrowsky, and M. Schwarz. 2007. Internet advertising and the generalized second-price auction: Selling billions of dollars worth of keywords. *American Economic Review* 97(1): 242–259.

Ellison, G., and S.F. Ellison. 2009. Search, obfuscation, and price elasticities on the internet. *Econometrica* 77(2): 427–452.

Ellison, G., and D. Fudenberg. 2003. Knife-edge or plateau: When do market models tip? *Quarterly Journal of Economics* 118(4): 1249–1278.

Ellison, G., D. Fudenberg, and M. Möbius. 2004. Competing auctions. *Journal of the European Economic Association* 2(1): 30–66.

Gabaix, X., and D. Laibson. 2006. Shrouded attributes, consumer myopia, and information suppression in competitive markets. *Quarterly Journal of Economics* 121(2): 505–540.

Hossain, T., and J. Morgan. 2006. . . .Plus shipping and handling: Revenue (non) equivalence in field experiments on eBay. *Advances in Economic Analysis & Policy* 6(2), Article 3.

Hossain, T., D. Minor, and J. Morgan. 2011. Competing matchmakers: An experimental analysis. *Management Science* 57(11): 1913–1925.

Iyer, G., D. Soberman, and J.M. Villas-Boas. 2005. The targeting of advertising. *Marketing Science* 24(3): 461–476.

Levin, J. D. 2011. The economics of internet markets. *National Bureau of Economic Research Working Paper Series,* No. 16852.

Narasimhan, C. 1988. Competitive promotional strategies. *Journal of Business* 61(4): 427–449.

Rochet, J.-C., and J. Tirole. 2003. Platform competition in two-sided markets. *Journal of the European Economic Association* 1(4): 990–1029.

Rochet, J.-C., and J. Tirole. 2006. Two-sided markets: A progress report. *RAND Journal of Economics* 37(3): 645–667.

Rosenthal, R.W. 1980. A model in which an increase in the number of sellers leads to a higher price. *Econometrica* 48(6): 1575–1579.

Shilony, Y. 1977. Mixed pricing in oligopoly. *Journal of Economic Theory* 14: 373–388.
Spulber, D. 1995. Bertrand competition when rivals' costs are unknown. *Journal of Industrial Economics* 43(1): 1–11.
Stigler, G. 1961. The economics of information. *Journal of Political Economy* 69(3): 213–225.
Tellis, G.J., E. Yin, and R. Niraj. 2009. Does quality win: Network effects versus quality in high tech markets. *Journal of Marketing Research* 46(2): 135–149.
Varian, H. 1980. A model of sales. *American Economic Review* 70(4): 651–659.
Varian, H. 2007. Position auctions. *International Journal of Industrial Organization* 25(6): 1163–1178.
Vickrey, W. 1961. Counterspeculation, auctions, and competitive sealed tenders. *Journal of Finance* 16: 8–37.
Weyl, E.G. 2010. A price theory of multi-sided platforms. *American Economic Review* 100(4): 1642–1672.

Pricing Services Online, Economics of

Anja Lambrecht

Firms deliver a variety of services online, ranging from content, software and banking to entertainment and networking. This article examines a firm's pricing decision for online services. It first discusses how a firm's decision of pricing services online differs from offline pricing decisions. It then discusses how firms can price services online. It examines the firm's choice between 'fee' or 'free' revenue models. It then turns to a firm's decision on its pricing structure. This includes the decision whether to sell or to rent, and the choice between pricing plans (e.g. pay-per-use, flat-rate tariffs or more complicated multi-part tariffs) or bundling. Lastly, it turns to the role of pricing in new product adoption.

Introduction

There is a large and growing literature in economics and marketing that relates to how firms price services that can be delivered online and how consumers respond to firms' pricing strategies. Broadly, this falls into two areas. First, researchers study a firm's choice between 'fee' or 'free' revenue models. This means that firms may offer the service for free and instead focus on other sources of revenue, such as advertising. Second, researchers examine the firm's choice of pricing structure. This includes the choice between selling or renting, or between different pricing plans or tariffs such as pay-per-use, flat-rate tariffs or more complicated multi-part tariffs. Lastly, the choice of pricing plan may play an important role in new product adoption, which is particularly important in markets with network effects where fast take-off is important for later success.

In this article, we refer to online services as services that are sold and delivered online, such as online content, networking or games. We discuss the firm's choice between different pricing strategies for such services. The Internet is also used to reach consumers to sell a large range of products or services that are delivered offline (e.g. amazon.com). For such goods, the Internet has enabled experimentation with a wide variety of pricing mechanisms that may be more difficult to implement offline. These mechanisms include online auctions (e.g. eBay), reverse selling (e.g. Priceline), advance purchasing (e.g. Groupon), and the use of investors as a revenue source (e.g. Kickstarter). Since these are not online services per se, but use the Internet merely as a channel for reaching customers, we do not consider such settings.

Services Online

Firms deliver a variety of services online, including content (e.g. huffingtonpost. com, nytimes. com), software (mcafeestore.com), banking (ingdirect. com), Internet connection (comcast. com), networking (linkedin.com, facebook. com) and entertainment (netflix.com, zynga.com, itunes.com).

A firm's choice of pricing services online differs from offline pricing decisions along three dimensions. First, while fixed costs of production or of providing the service infrastructure may be substantial, the marginal cost of serving an additional customer online is most often zero or close to zero. This gives the firm great flexibility in pricing. Second, many firms may offer the service free of charge and instead sell advertising. The ability to select among different streams of revenue has both broadened and complicated a firm's choice of revenue models, which is no longer restricted to setting a price level. As a result, a firm may face a competitor that offers their service for free. Third, in digital environments it is relatively easy to meter a consumer's usage accurately and to implement usage-based price discrimination.

Fee or Free Pricing

Many firms that sell services online can choose either to charge users for access to all or part of their service, or to provide the service for free. When services are provided for free, firms sell advertising or other complementary services. Online content providers, for example, have chosen a variety of fee or free pricing strategies. For example, the *Los Angeles Times* offers all content for free, the *Times* requires a subscription to view any articles, and the *New York Times* offers 10 articles a month for free, but only subscribers can view additional articles. Broadly, the economics and marketing literature has taken two perspectives on a firm's choice between free and fee strategies.

Sampling when firms use sampling, they offer a free sample of the product and require consumers to pay for full usage (e.g. nytimes.com). The idea is that after sampling a service, consumers become more likely to sign up for its full version, which increases long-term sales (Bawa and Shoemaker 2004). Additionally, for digital goods, free samples alongside high prices can signal superior quality (Boom 2010). Halbheer et al. (2013) show that the choice of a sampling strategy is determined by the relationship between advertising effectiveness and content quality. Offering only paid content is optimal under low advertising effectiveness. For intermediate levels of advertising effectiveness, the publisher should switch to a sampling strategy. Only under high levels of advertising effectiveness is it optimal to offer all content for free.

Trading Off Advertising and Subscription Revenues the basic trade-off for firms that provide services for free is that charging for access to content reduces the number of page views by users and hence the potential for advertising revenues. The challenge is to identify when a 'free' or a 'fee' strategy may be optimal.

In early research, Shapiro and Varian (1998) and Bhargava and Choudhary (2001) show that offering both a paid and a free component can allow firms to implement quality differentiation, versioning or seconddegree price discrimination. Godes et al. (2009) more explicitly examine the trade-off between subscription and advertising revenues. They find that since greater competitive intensity may reduce advertising revenues, a firm in a duopoly is less willing to under-price content to increase demand than a monopolist. As a result, greater competitive intensity may increase profits from charging for content and decrease profits from advertising. However, offering paid content can lead to both a loss in visitors and to a positional disadvantage in advertising markets, since advertisers are willing to pay a premium to firms with a high expected share of loyal consumers (Athey et al. 2013). Prasad et al. (2003) do not consider the competitive setting, and instead focus on the effect of consumer heterogeneity in their willingness to pay to avoid ads. They find that in most cases the firm should combine payper- view and advertising revenues, rather than relying exclusively on either pay-per-view or advertising revenues. As Halbheer et al. (2013) illustrate, advertising effectiveness and content quality will further determine whether the firm should charge for access to content. These results illustrate the complexity of a firm's decision that needs to account for the competitive setting in the market for consumers and the market for advertisers, the attractiveness of its content, and the heterogeneity in consumers' tastes.

Empirically, Pauwels and Weiss (2007) find for an online content provider targeted towards marketing professionals that moving from free to fee can be profitable, despite the loss of advertising revenue. However, it is not clear whether such insights readily apply to consumer markets. Indeed, recent pricing research finds that among consumers the demand effect of changing from a zero to a small non-zero price may be significantly more pronounced than what price elasticities evaluated at other points of the demand curve would suggest (Odlyzko 2001; Shampanier et al. 2007; Ascarza et al. 2012). This likewise suggests that in consumer markets relatively low fees for online services may strongly discourage consumers from visiting the site.

In line with this insight, Chiou and Tucker (2011) find a strong negative effect of the introduction of a paywall by an online news site, particularly among younger consumers. Using micro-level data from the sports site ESPN, Lambrecht and Misra (2012) quantify the tradeoff between greater subscription and lower advertising revenues from offering paid content and find that whether the firm benefits from adding an additional paid article varies by whether a sport is offseason, in regular season or in post-season. They attribute these differences to a variation in the value of sport news across seasons and suggest that firms should pay attention to how consumer valuation of online content varies across time and dynamically adjust the amount of paid content. (Note that a firm's decision to sell advertising and offer the service for free may also affect the type of content a firm provides. Sun and Zhu (2012) show that when incentivised by ad revenues, blogs are more likely to show more popular content.)

In sum, these insights document that whether a firm can successfully charge for access to its service is related to consumers' valuation of the service, which may vary across customers and across time. They also show that in evaluating whether a fee or a free strategy is optimal, the firm needs to quantify both additional revenues from subscriptions and the loss in advertising revenues due to the increase in page views.

Note that how valuable a site is for advertisers also depends on how well such a site can target consumers. Search engines, for example, have proven highly valuable for advertising. By typing search terms, consumers directly reveal their intentions and preferences, including in many instances their intentions to purchase (Edelman 2009). As a result, Google is able to generate almost all its profits from advertising revenues and can provide the service for free to searchers. Other sites that offer neither a high-quality online environment that may allow brand advertising nor are able to target consumers may only be able to generate low advertising revenues. (Recently developed tracking techniques also allow firms that sell display advertising to target consumers with ads suited to their specific interests and purchase intentions (Lambrecht and Tucker 2013)).

Setting the Price Structure

A firm that charges for access to services needs to determine how to price its offering optimally. The ability to meter consumers' usage behaviour in real time on the level of each individual consumer means that firms may implement a wide range of pricing formats, including selling vs. renting, charging flat fees versus charging for usage, and different forms of usage-based price discrimination.

Selling vs. Renting Internet technology allows many firms to choose between renting or selling digital content to consumers. Renting usually limits a consumer's right to use a service (e.g. watch a movie) to a fixed period of time, whereas purchase grants unlimited rights of usage. Rao (2011) analyses whether an online movie provider should focus on a sell or rent strategy. She finds that consumer heterogeneity in one-time versus repeat consumption preferences drives purchase and rental offerings. As a result, a firm can use purchase and rental markets to differentiate between consumers.

Even when consumers typically require repeated usage, attempts have been made to rent products as a service instead of selling physical products. The firm is able to 'servify' the product, since digital technology allows close monitoring and charging for an individual's usage of a product. Such attempts to turn physical products into services go back to the early days of the Internet. For example, Electrolux Sweden piloted installing washing machines for free in consumers' homes and asked consumers to payper- wash, providing the service 'clean clothes' instead of a physical product (http://group.electrolux.com/en/electrolux-offers-7000-householdsfree- washing-machines-1885/). However, despite initial enthusiasm about the ability to price discriminate between consumers who require little usage (and hence would be better off renting the service) and intense users (who benefit from purchasing), such pricing techniques have never been fully embraced. Research suggests that this can be linked to consumers' preferences for paying flat fees rather than per use.

Flat Fees or Pay-Per-Use when a firm is able to meter consumers' usage it can charge for actual usage instead of a flat fee, so a consumer's bill more accurately reflects their consumption (Levinson and Odlyzko 2008). Yet research finds that consumers often prefer a flat-rate tariff to a tariff that charges for actual usage, even if the consumer's bill ex post would be lower on a tariff that charges per usage (Train et al. 1987; Kridel et al. 1993; Lambrecht and Skiera 2006; DellaVigna and Malmendier 2006). Lambrecht and Skiera (2006) document three reasons for the so-called flat-rate bias. First, consumers' dislike of metered usage may lead them to enjoy their usage less than they otherwise would (see also Prelec and Loewenstein 1998), also referred to as the 'taxi meter effect'. Second, consumers prefer a steady bill to variation in their bill over time, which can be linked to risk aversion and loss aversion. Third, consumers may over-estimate their usage and mistakenly believe they have chosen the optimal tariff. Findings by Iyengar et al. (2011) further confirm consumers preferences for flat-rate tariffs. Their research shows that customers derive a lower utility of usage under a tariff that charges per usage than under a flat-rate tariff.

Consumers' preferences for flat-rate versus pay-per-use pricing represent both advantages and challenges. It means that consumers often pay more than they would on a pay-per-use tariff, increasing firm revenues and consumer lifetime value (Lambrecht and Skiera 2006). But under flat-rate tariffs, a small proportion of very high-usage consumers may make a service offering unprofitable. Even under zero marginal costs, a firm may face capacity constraints (e.g. of the Internet access network or the amount of traffic a website can handle). To deal with particularly high users, some Internet service providers have started to impose usage caps, effectively cutting off consumers once they exceed a set limit (Edelman 2009). Alternatively, firms in various sectors have turned to usage-based price discrimination.

Usage-Based Price Discrimination two-part tariffs are the classic way to price-discriminate based on consumers' usage. A two-part tariff charges an access price and a usage-price for each consumed unit. When a firm offers a menu of optimal two-part tariffs, consumers self-select into a tariff, allowing the firm to price-discriminate between consumers (Oi 1971; Schmalensee 1981; Tirole 1988; Wilson 1993; for a summary see also Lambrecht et al. 2012). However, since consumers tend to dislike the pay-per-use structure of two-part tariffs, managerial practice has since moved to more complex pricing structures.

Specifically, three-part tariffs and bucket pricing have become increasingly prominent (Jensen 2006; Bagh and Bhargava 2013; Iyengar et al. 2007; Lambrecht et al. 2007; Grubb and Osborne 2012; Schlereth and Skiera 2012). Both three-part tariffs and bucket pricing charge an access price and offer an allowance of free units of consumption (e.g. free articles of an online content provider or free minutes by a mobile telephony service provider). Within this allowance, consumers are not charged a usage price and the consumer's cost function thus has the structure of a flat-rate tariff.

For any usage in excess of the usage allowance, a three-part tariff charges per unit of consumption, similar to a two-part tariff. Research has shown that, similar to the flat-rate bias, consumers choosing among multiple threepart tariffs that differ in the size of their usage allowance and access fees tend to choose a tariff with a higher allowance than optimal based on their ex post usage (Lambrecht and Skiera 2006).

Such behaviour is a result of their tariff-specific preferences, but also a rational outcome of their two-step decision process. A consumer initially chooses a tariff based on their expected usage, but is uncertain about the exact value of their ex-post consumption. Later, the consumer decides on their usage conditional on their previous tariff choice. Lambrecht et al. (2007) show that as usage uncertainty increases, tariffs with increasingly higher allowances and access fees become the optimal choice. As a result, usage uncertainty increases a provider's profits under three-part tariffs, but hurts consumer surplus. Evidence from Ascarza et al. (2012) illustrates that three-part tariffs may affect not only tariff choice but also usage. They find that consumers who switch from a two-part tariff to a three-part tariff that provides a 'free' component have a higher valuation of usage. As a result, they use more on the three-part tariff than would be expected based on their previous two-part tariff usage. This pattern reflects usage behaviour when AOL replaced their two-part tariffs with flat-rate tariffs in 1996 – over the next year usage tripled (Odlyzko 2001).

Bucket pricing limits consumption to the allowance or 'bucket' of free units. It does not allow incremental usage above the allowance, though some bucket pricing plans instead offer the possibility to purchase additional allowances of units (Schlereth and Skiera 2012). Web hosting firms, for example, may offer bucket pricing where a bucket includes an allowance of web space and, potentially, domains and applications (for example 1and1.co. uk/hosting). Similarly, mobile phone providers may offer a bucket that includes an allowance of minutes call volume and an allowance of data transfer volume (vodafone. co.uk).

Firms often choose bucket pricing, instead of three-part tariffs, when a bucket is defined by multiple attributes. It would then be more difficult to set up a three-part tariff where the bill would increase with usage along one dimension only and to communicate such a structure to consumers. Alternatively, some service attributes may be discrete rather than continuous choices (such as adding different applications), which may justify in the consumer's mind a step change in price.

Bundling online, it is very easy for firms to bundle since the marginal cost of 'repackaging' individual goods as bundles is low. Examples for bundling online includes songs bundled as a virtual CD or subscriptions for software packages such as Microsoft Office.

Bundling in a monopoly is generally profitable under two conditions: marginal costs are low and demand is heterogeneous (Crawford 2008; Bakos and Brynjolfsson 1999; Fang and Norman 2006; Olderog and Skiera 2000). For many services that can be provided online both are indeed the case. For example, the marginal cost of selling an additional song is very low. Likewise, consumers typically have heterogeneous tastes for music. Bakos and Brynjolfsson (2000) show how, even under competition, bundling can be optimal. This includes both upstream and downstream competition, competition between a bundler and a single good and between two bundlers.

Yet bundling of services online is not as widespread as one might expect. As discussed earlier, news stories are still largely offered for free instead of being sold in a subscription package. Online, songs are purchased more often separately than bundled as a virtual CD. The reason is the high degree of competition online alongside the possibility for firms to provide content for free (and instead sell advertising). Even competitive provision of only some of the components of the bundle, such as individual songs or news stories, significantly lowers consumers' willingness to pay for the bundle. This further illustrates the difficulty of a 'fee' strategy for online content and explains why only clearly differentiated online news sites such as the *Wall Street Journal, Financial Times* and *New York Times* have moved to a 'fee' model. It suggests that only firms that are able to clearly differentiate their service and brand will be able to bundle their services online.

Price as Barrier to Adoption

Consumers who consider adopting new online services often face a considerable cost of doing so. First, consumers often have high uncertainty about how much they value a new product.

Second, adopting a new service often requires multiple types of switching costs (Klemperer 1987). This may include monetary costs. It may also include non-monetary costs, such as the cost of time and effort to set up or learn how to operate a new service. Since consumers are sensitive to even small monetary prices (Shampanier et al. 2007), any positive price may inhibit product adoption by consumers who initially have high uncertainty about how much they value a new service. This is particularly challenging for firms in markets with network effects. When quickly reaching a large customer base determines long-term success, firms may opt to offer a service free of charge and instead charge advertising revenues (e.g. facebook.com), use sampling or charge a premium for advanced features, functionalities or virtual goods. The latter are sometimes referred to as 'freemium' models. For example, online gamers can often use the basic version of the game for free, but higher levels may require payment. Alternatively, users can buy add-ons or tokens for use in the game, e.g. to progress quicker to higher levels or for additional functionalities (zynga.com).

The benefit of 'freemium' is that a zero price eliminates one barrier to adoption and the firm can more easily acquire customers with initially high uncertainty about how much they value the good. With increasing product experience, a consumer's uncertainty decreases and, on average, their valuation increases. When firms require only more advanced consumers to pay, they can price-discriminate between new consumers with, on average, low valuation and high uncertainty and experienced consumers with, on average, high valuation and low uncertainty. Additionally, they can exploit consumer lock-in that may arise from learning how to use the product and increased product familiarity.

Non-monetary cost of time and effort, such as the hassle of installing a new service, can also impose significant costs on a consumer, and inhibit new services adoption. For example, when choosing a web hosting provider consumers may expect high hassle costs, in particular how easy it will be to set up a website with a specific hosting provider. Such hassle costs may deter consumers. Lambrecht and Tucker (2012) show how firms that sell service contracts can set their prices to attenuate the negative effect of hassle costs on adoption. Specifically, they should discount the period for which consumers expect hassle costs, even if this means slightly increasing prices in other periods.

Summary and Conclusion

This article examines how firms price services online, covering the decision whether to charge (fee or free), and then how to charge (rent or sell, pay-peruse, flat-rate, multi-part tariffs, bundling). Lastly, it considers how to price new services, in the context of overcoming price barriers to adoption. Reviewing the recent academic literature, the article examines the role played by usage caps and complex pricing structures in creating sustained profits, and emphasises the importance of close attention to consumer data and the consumer experience in setting up pricing structures.

See Also

- ▶ Google
- ▶ Pricing on the Internet
- ▶ Online Platforms, Economics of

Acknowledgment The author thanks Gregory Crawford and Bernd Skiera for their comments.

Bibliography

Ascarza, E., A. Lambrecht, and N.J. Vilcassim. 2012. When talk is 'free': An analysis of subscriber behavior under two- and three-part tariffs. *Journal of Marketing Research* 49(6): 882–899.

Athey, S., E. Calvano, and J.S. Gans. 2013. *The impact of the Internet on advertising markets for news media*, Working paper. Cambridge, MA: National Bureau of Economic Research.

Bagh, A., and H.K. Bhargava. 2013. How to price discriminate when tariff size matters. *Marketing Science* 32: 111–126.

Bakos, Y., and E. Brynjolfsson. 1999. Bundling information goods: Pricing, profits and efficiency. *Management Science* 45(12): 1613–1630.

Bakos, Y., and E. Brynjolfsson. 2000. Bundling and competition on the Internet. *Marketing Science* 19(1): 63–82.

Bawa, K., and R. Shoemaker. 2004. The effects of free sample promotions on incremental brand sales. *Marketing Science* 23(3): 345–363.

Bhargava, H.K., and V. Choudhary. 2001. Information goods and vertical differentiation. *Journal of Management Information Systems* 18(2): 89–106.

Boom, A. 2010. *'Download for free'- when do providers of digital goods offer free samples?*. Working paper.

Chiou, L. and C. Tucker. 2011. *Paywalls and the demand for news*. Working paper.

Crawford, G. 2008. The discriminatory incentives to bundle: The case of cable television. *Quantitative Marketing & Economics* 6(1): 41–78.

DellaVigna, S., and U. Malmendier. 2006. Paying not to go to the gym. *American Economic Review* 96: 694–719.

Edelman, B. 2009. Priced and unpriced online markets. *Journal of Economic Perspectives* 23(3): 21–36.

Fang, H., and P. Norman. 2006. To bundle or not to bundle. *Rand Journal of Economics* 37(4): 946–963.

Godes, D., E. Ofek, and M. Sarvary. 2009. Content vs. advertising: The impact of competition on media firm strategy. *Marketing Science* 28(1): 20–35.

Grubb, M. and M. Osborne. 2012. *Cellular service demand: Biased beliefs, learning, and bill shock. MIT Sloan Research Paper No. 4974– 12.* Chestnut Hill: Boston College, Department of Economics.

Halbheer, D., F. Stahl, O. Koenigsberg, and D.R. Lehmann. 2013. *Digital content strategies. Columbia Business School Research Paper No. 11–8*. Zürich Univ., Inst. für Betriebswirtschaftslehre.

Iyengar, R., A. Ansari, and S. Gupta. 2007. A model of consumer learning for service quality and usage. *Journal of Marketing Research* 44(4): 529–544.

Iyengar, R., K. Jedidi, S. Essegaier, and P.J. Danaher. 2011. The impact of tariff structure on customer retention, usage and profitability of access services. *Marketing Science* 30(5): 820–836.

Jensen, S. 2006. Implementation of competitive nonlinear pricing: Tariffs with inclusive consumption. *Review of Economic Design* 10: 9–29.

Klemperer, P. 1987. Markets with consumer switching costs. *Quarterly Journal of Economics* 102(2): 375–394.

Kridel, D.J., D.E. Lehman, and D.L. Weisman. 1993. Option value, telecommunication demand, and policy. *Information Economics and Policy* 5: 125–144.

Lambrecht, A. and Misra, K. 2012. *Pricing online content: Fee or free?*. Working paper.

Lambrecht, A., and B. Skiera. 2006. Paying too much and being happy about it: Existence, causes and consequences of tariff-choice biases. *Journal of Marketing Research* 43(May): 212–223.

Lambrecht, A., and C. Tucker. 2012. Paying with money or with effort: Pricing when customers anticipate hassle. *Journal of Marketing Research* 49: 66–82.

Lambrecht, A. and C. Tucker. 2013. *When does retargeting work? Information specificity in online advertising*. Working paper.

Lambrecht, A., K. Seim, and B. Skiera. 2007. Does uncertainty matter? Consumer behavior under three-part tariffs. *Marketing Science* 26(5): 698–710.

Lambrecht, A., K. Seim, N. Vilcassim, A. Cheema, Y. Chen, G.S. Crawford, K. Hosanagar, R. Iyengar, O. Koenigsberg, R. Lee, E.J. Miravete, and O. Sahin. 2012. Price discrimination in service industries. *Marketing Letters* 23: 423–438.

Levinson, D., and A. Odlyzko. 2008. Too expensive to meter: The influence of transaction costs in transportation and communication. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 366(1872): 2033–2046.

Odlyzko, A. 2001. Internet pricing and the history of communications. *Computer Networks* 36: 493–517.

Oi, W.Y. 1971. A Disneyland dilemma: Two-part tariffs for a Mickey Mouse monopoly. *Quarterly Journal of Economics* 85: 77–96.

Olderog, T., and B. Skiera. 2000. The benefits of bundling strategies. *Schmalenbach Business Review* 1(2): 137–160.

Pauwels, K. and A. Weiss. 2007. Moving from free to fee: How marketing can stimulate gains and stem losses for an online content provider. *Journal of Marketing*, 72(3)

Prasad, A., V. Mahajan, and B. Bronnenberg. 2003. Advertising versus pay-per-view in electronic media. *International Journal of Research in Marketing* 20: 13–30.

Prelec, D., and G. Loewenstein. 1998. The red and the black: Mental accounting of savings and debt. *Marketing Science* 17(1): 4–28.

Rao, A. 2011. *Online content pricing: Purchase and rental markets*. Working paper.

Schlereth, C., and B. Skiera. 2012. Measurement of consumer preferences for bucket pricing plans with different service attributes. *International Journal of Research in Marketing* 29(2): 167–180.

Schmalensee, R. 1981. Monopolistic two-part tariff arrangements. *Bell Journal of Economics* 25: 445–466.

Shampanier, K., N. Mazar, and D. Ariely. 2007. Zero as a special price: The true value of free products. *Marketing Science* 26(6): 745–757.

Shapiro, C., and H.H. Varian. 1998. Versioning: The smart way to sell information. *Harvard Business Review* 76: 106–114.

Sun, M. and F. Zhu. 2012. Ad revenue and content commercialization: Evidence from blogs. *Management Science*, 59: 2314–2331, Published online ahead of print April 4, 2013.

Tirole, J. 1988. *The theory of industrial organization*. Cambridge, MA: MIT Press.

Train, K.E., D.L. McFadden, and M. Ben-Akiva. 1987. The demand for local telephone service: A fully discrete model of residential calling patterns and service choices. *Rand Journal of Economics* 18(1): 109–123.

Wilson, R. 1993. *Nonlinear pricing*. New York: Oxford University Press.

Primary and Secondary Labour Markets

Paul Ryan

Keywords

Dual economies; Duality; Labour market discrimination; Primary and secondary labour markets; Productivity growth; Segmentationism; Uneven development; Wage dispersion

JEL Classifications

J42

The primary/secondary distinction involves an application of the concept of economic dualism to the labour markets of advanced capitalist economies.

In the initial formulation, the primary and secondary segments of a dual labour market were distinguished principally by job characteristics. The rewards of primary jobs, in terms of earnings, working conditions, job security, training opportunities and career prospects, are high; those of secondary jobs, low. Increases in a worker's schooling and work experience lead to higher job rewards in the primary segment but not in the secondary one. Inter-segment mobility is limited, the working poor being confined to secondary jobs. A separate dichotomy in worker traits parallels that in jobs. Secondary workers are those with weak attachment to employment, a consequence of social roles in either the household (youths and married females) or the locality (inner-city minorities; Piore 1970).

Two important differences soon emerged in dualist interpretation. The central difference between the segments for some authors involves stability of employment; for others, pay levels (Piore 1970; Bluestone 1970). Some see the dualist classification as partial; others seek to classify all jobs and workers within an exhaustive schema. Exhaustiveness has become predominant, with the ensuing heterogeneity of an enlarged primary segment leading to further dualisms (upper/lower tier and core/periphery, by occupation and industry respectively) within primary employment (Bluestone 1970; Edwards et al. 1975; Piore 1975).

Dualist interpretations originate from two sources. The first is the failure in the 1960s of a manpower policy oriented to the enhancement of individuals' job skills to move large numbers of US inner-city residents into stable and well-paid work. The explanation was sought in the characteristics not of workers but of jobs, with the primary/secondary duality building upon the antecedent structured/unstructured one (Kerr 1954). The second source is the concern of radical economists to understand the political disunity of the US labour force, a painful anomaly for Marxist analysis. The key to political fragmentation has been sought in the differentiation of work experiences in a dual labour market (Edwards 1979; Gordon et al. 1982).

Dualism is a variant of segmentationism, sharing with it three attributes which distinguish both from orthodox labour theory. The first is the widening of the analytical scope beyond comparative statics with given preferences and indeterminate public policy. Thus the instability of inner-city employment is attributed to an interaction between worker attitudes and job attributes, with attitudes thereby made endogenous. The secondary status of female and youth labour is understood in terms not of autonomous preferences but rather of power relations within family and state (Humphries and Rubery 1984).

Secondly, the labour market is seen as systematically differentiating the job rewards achieved by comparable individuals. The market then becomes a source of inequality in its own right. Thus dualism in employment stability is understood to result not so much from the aversion of secondary workers to steady work as from their discriminatory exclusion from stable jobs. Similarly, the low pay of secondary workers is explained not so much in terms of low labour quality as of denial of access to the primary jobs which convert high potential into high actual productivity (Ryan 1981).

Finally, labour market outcomes such as pay and turnover are seen as determined principally by such product market attributes as demand variability, employer power and production technology. The part played by labour market influences, including trade unions, is a subsidiary one. An important role is given to competitive forces in determining labour outcomes, but such forces derive more from the product than from the labour market (Wilkinson 1981).

These three attributes rebut the criticism that the dualist and segmentationist approach is largely descriptive, taxonomic and compatible with competitive theory (Wachter 1974; Cain 1976).

Considering dualism as a subset of segmentationism, two interpretations may be placed upon their relationship. The first is descriptive. *Heuristic duality* describes vividly the idea of differential treatment in labour markets without implying discontinuity or universality. Thus to distinguish good and bad jobs for comparable workers need not rule out large numbers of medium jobs and unclassified jobs. Heuristic dualism is also seen in the distinction between sheltered and exposed sectors, familiar in 1920s Britain, when currency overvaluation depressed relative wages according to exposure to foreign competition (Dobb 1928). To postulate a sheltered/exposed dualism is to dramatize the issue without requiring that exposure itself be dichotomous or that a comprehensive theory of labour outcomes be built on such a limited basis.

The second interpretation of dualism is more demanding. *Strict duality* requires not just a substantial dispersion of job rewards for comparable individuals but also the polarization of their distribution into two clearly separate segments, each with low internal heterogeneity; a substantial distance between average job rewards in the two segments; and few cases falling in the intermediate range. Such bimodality must maintain and reproduce itself over time, while individuals and jobs should show low rates of mobility across a clear intervening boundary. Such conditions may fail to be realized literally in practice but strong tendencies towards them are required for strict dualism to be sustained.

Although the heuristic and the strict formulations of duality are frequently confused, leading dualist writers have explicitly espoused strict duality. The causes of a postulated strict dualism in job rewards have been sought in underlying dichotomies in three dimensions of industrial structure. The first explanation sees labour dualism in terms of employment stability. Selective worker organization in pursuit of job security leads to primary jobs in firms producing for the stable portion of product demand, with unstable secondary jobs where employers sell to the variable or unpredictable part (Berger and Piore 1980). The second approach sees dualism in terms of earnings, relating it to an underlying dichotomy amongst firms and industries in market and political power (Averitt 1968). The third explanation distinguishes firms whose organizational structures motivate and control their employees by providing stable jobs, career prospects and high pay from those which rely upon the traditional methods of low pay, insecurity and discipline. This dichotomy in control techniques overlaps with the preceding one by producer power, it being the large and powerful corporations which adopt sophisticated control strategies (Gordon et al. 1982).

These three theories of strict dualism all capture important sources of segmentation in labour markets. An empirical role is most evident for producer power, in the shape of significant associations between employee rewards and such power correlates as seller concentration, firm size and ties to the state. However, strict dualism oversimplifies the links between industrial structure and labour outcomes. These theories offer no reason for the distributions of demand variability, producer power or control strategies to become polarized in the first place. In practice, the nexus between product and labour markets proves empirically multidimensional and complex (Wallace and Kalleberg 1981; Hodson and Kaufman 1982). Moreover, while bimodality has been found in some attributes of industrial structure, this typically appears in only one of a set of several attributes; is found in data-sets which exclude more than half of national employment; and even then does not lead to any strict dualism in labour outcomes (Oster 1979; Buchele 1983).

The empirical status of segmentation (and heuristic duality) remains controversial, reflecting the difficulty of measuring labour quality and market structure. The evidence concerning strict duality, is, however, distinctly unfavourable. No clear boundary emerges between segments. Definitions of the secondary segment vary widely in size and composition from one study to another, with intermediate groups proving numerous and difficult to classify. The difference in average job rewards between segments in most dualist definitions proves only moderate in earnings and erratic in employment stability. Mobility between segments appears too high to support an inference of wholesale confinement to secondary employment.

The empirical failure of strict dualism in the domestic economy may be understood by considering a more promising candidate: the world labour market, treated as a potential whole (Singer 1970). The gap between the earnings of comparable workers in the two poles of advanced and developing countries is great; intermediate cases (the newly industrializing countries) are certainly numerous, but bimodality is still expected; while the distance in earnings between the two poles has proved not only durable but at times even increasing, with the attainment of higher rates of growth in productivity and earnings in advanced than in developing countries (Brandt Commission 1980).

Similar forces for dualist divergence function within the labour markets of both the advanced economies and the world economy. The international phenomena of uneven development and cumulative divergence, resting upon the attainment of higher rates of investment and productivity growth in advanced then in developing countries, have as their national counterpart the large and persistent differences in productivity growth across sectors (Salter 1960). Unequal exchange, or the systematic overvaluation of the output of advantaged countries at the expense of that of weaker ones, also finds its domestic analogue in the output prices of primary and secondary segment employers.

One reason why strict dualism applies more to the international than to the national labour market involves the greater obstacles to factor mobility across than within national boundaries. Two other influences are potentially more important. First, the dispersion of rates of growth of value productivity within advanced economies is limited relative to its international counterpart. In the domestic economy, sectors with low rates of growth of physical productivity experience either the transfer of production to developing countries (in the case of tradables) or the revaluation of their output by increases in relative price (in the case of non-tradables). The former mechanism has no counterpart in the international context. Second, the world economy lacks the institutions which prevent differences in rates of growth of value productivity from producing increasing dispersion (let alone polarization) within the distributions of job rewards of the advanced economy. Relativity bargaining (for the organized), statutory wage minima and indexed social security provision (for the unorganized) prevent substantial widening of the gap between earnings in high and low productivity growth employment. The only counterpart to these forces in the world economy is development aid, a pale reflection of social security in the domestic economy. The polarization of labour outcomes is therefore possible in the world labour market to an extent inconceivable in the domestic one.

The factors which curb the dispersion of labour outcomes within advanced economies have been weakened lately by the growth of unemployment and anti-regulatory sentiment. They remain nevertheless sufficiently powerful to restrain domestic tendencies toward dualist divergent development. The dual labour market provides a tenable account of labour market segmentation within advanced economies only in its weaker, heuristic formulation.

See Also

- ▶ Labour Market Institutions

Bibliography

Averitt, R.T. 1968. *The dual economy*. New York: Norton.
Berger, S., and M. Piore. 1980. *Dualism and discontinuity in industrial societies*. Cambridge: Cambridge University Press.
Bluestone, B. 1970. The tripartite economy. *Poverty and Human Resources Abstracts* 5 (4): 15–35.
Brandt Commission. (Independent Commission on International Development Issues). 1980. *North–South: A programme for survival*. London: Pan Books.
Buchele, R.K. 1983. Economic dualism and employment stability. *Industrial Relations* 22: 410–418.
Cain, G.C. 1976. The challenge of segmented labor market theories to orthodox theory: A survey. *Journal of Economic Literature* 14: 1215–1257.
Dobb, M. 1928. *Wages*. London: Nisbet.
Doeringer, P.B., and M. Piore. 1971. *Internal labor markets and manpower analysis*. Lexington: D.C. Heath.
Edwards, R.C. 1979. *Contested terrain*. New York: Basic Books.
Edwards, R.C., M. Reich, and D.M. Gordon, eds. 1975. *Labor market segmentation*. Lexington: D.C. Heath.
Gordon, D.M., R.C. Edwards, and M. Reich. 1982. *Segmented work, divided workers*. Cambridge: Cambridge University Press.
Hodson, R., and R. Kaufman. 1982. Economic dualism: A critical review. *American Sociological Review* 47: 727–739.
Humphries, J., and J. Rubery. 1984. The reconstitution of the supply side of the labour market. *Cambridge Journal of Economics* 8: 331–346.
Kerr, C. 1954. The balkanization of labor markets. In *Labor mobility and economic opportunity*, ed. E.W. Bakke. Cambridge, MA: MIT Press.
Oster, G. 1979. A factor analytic test of the theory of the dual economy. *Review of Economics and Statistics* 61: 33–39.
Piore, M. 1970. The dual labor market; Theory and implications. In *The state and the poor*, ed. R. Barringer and S.H. Beer. Cambridge, MA: Winthrop.
Piore, M. 1975. Notes for a theory of labour market stratification. In *Labor market segmentation*, ed. Reich Edwards and Gordon. Lexington: Heath.
Ryan, P. 1981. Segmentation, duality and the internal labour market. In *The dynamics of labour market segmentation*, ed. F. Wilkinson. London/New York: Academic.
Salter, W.E.G. 1960. *Productivity and technical change*. Cambridge: Cambridge University Press.
Singer, H.W. 1970. Dualism revisited: A new approach to the problems of the dual society in developing countries. *Journal of Development Studies* 7 (1): 60–75.
Wachter, M. 1974. Primary and secondary labor markets: A critique of the dual approach. *Brookings Papers on Economic Activity* 1974 (3): 637–680.
Wallace, M., and A. Kalleberg. 1981. Economic organization of firms and labor market consequences: Towards a specification of dual economic theory. In *Sociological perspectives on labor markets*, ed. I. Berg. New York: Academic.
Wilkinson, F., ed. 1981. *Dynamics of labor market segmentation*. New York: Academic.

Primitive Capitalist Accumulation

Ross Thomson

The primitive (or original) accumulation of capital is a concept developed in Karl Marx's *Capital* and *Grundrisse* to designate that process which generates the preconditions of the ongoing accumulation of capital. The character of these preconditions is derived from the concept of capital, understood to be the process whereby money is invested in the purchase of means of production and labour-power (the worker's capacity to labour) which in turn produce commodities embodying surplus-value. Capital therefore presupposes money amassed to be accumulated, labour-power as the property of labourers separated from ownership of the means of production, and markets in which commodities can be sold. Primitive accumulation therefore must involve more than Adam Smith's notion that 'The accumulation of stock must, in the nature of things, be previous to the division of labour' (1776, p. 260), whether the stock consists of money, means of production or means of subsistence. For this notion ignores the need for a proletariat, the importance of which is shown by settler colonies which have wealth but, insofar as the availability of land precludes the emergence of a market for labour-power, no capital.

To grasp the process generating the preconditions of capital entails historical investigation, which for Marx focused principally on the first industrial capitalist power, England, during the historical period, extending from the mid-sixteenth century through 1770, called the stage of manufacturing. Primitive accumulation consisted of several distinct processes which transformed each of the elements of the inherited division of labour between the towns and the countryside: landed property which combined common with private rights of landlords and free peasant proprietors, merchant capital in wholesale trade, and craft production centred in the urban trades. We will identify and evaluate Marx's account of these processes and will then consider whether this account helps understand the rise of English industrial capitalism and the processes of primitive accumulation elsewhere. Partly to remedy misunderstandings brought about by Marx's intentionally one-sided emphasis on the role of force, we will emphasize the economic mechanisms at work.

The Agricultural Revolution

For Marx (1890, chs 27, 29), the first and foremost effect of the 'agricultural revolution' of the sixteenth through eighteenth centuries was to expropriate the peasant from the soil and establish capitalist agriculture. Marx argues that a new, money-oriented nobility and gentry forcibly enclosed desmesne, common and waste land, consolidated small farms into larger ones and at times converted to pasturage. Capitalist farmers grew from a differentiation of the peasantry. By 1800, both yeoman and communal rights had been eliminated.

While Marx did overemphasize both the coerciveness and the significance of enclosures, his basic point that a landless proletariat and capitalist agriculture had become widespread in the manufacturing period remains valid. Enclosures converted property characterized by shared rights into private property. Although enclosures usually accorded with the custom of the manor and were undertaken by agreement of those with property rights, they did rely on the local power of landlords and, especially in the second half of the eighteenth century, the centralized power of the state. As Tawney emphasized (1912), they were an important means of expropriation of those without legally enforceable rights to their land, notably leaseholders, squatters and cottagers.

But other factors may have been more important in separating peasants from the land. Engrossment combined many small farms into few larger farms and therefore replaced small leaseholders by larger capitalist tenants. The differentiation of the peasantry led to land sales by some (Lenin 1908; Dobb 1947). This process was facilitated by the presence of a land market and the growth of population from 1500 to 1640 and again after 1750. Demographic expansion among the landless further increased the numbers of proletarians (Tilly 1984).

Marx (1890, ch. 30) maintained that the transformation of agriculture had the significance of creating a proletariat for industry as well as agriculture. The supply of both agricultural goods and labour-power for other sectors of the economy increased as a result of growing labour productivity, a second facet of the agricultural revolution, combined with more intense work and lower consumption by workers compared to smallholders. This argument has received support from recent agrarian history, which points to productivity growth coming from convertible husbandry, new rotations including grasses and the turnip, and greatly improved animal husbandry (Chambers and Mingay 1966; Kerridge 1967; Jones 1974). Such innovation may have been aided by the accumulation of capitalist farmers and by the control and scale afforded by enclosure and engrossment. Moreover, enclosures were often depopulating, especially when they led to convertible husbandry or pasturage. Such changes allowed the share of nonagricultural population to rise from 40 per cent in 1688 to 64 per cent in 1801 in a period when England was largely self-sufficient in foodstuffs.

Finally, Marx correctly contended that with the decline of subsistence production, wage-labourers contributed to the expansion of the home market. But especially in periods of rising prices like the sixteenth century, the growing rural middle class may have added even more to market expansion, particularly for industrial products. Growing productivity may also have supported the home market by causing relative agricultural prices to fall, so that incomes in the industrial sector could rise while the income of farmers need not decline (John 1965; Jones 1974).

Commercial Accumulation and Market Expansion

The genesis of capitalist agriculture contrasts sharply with the birth of capitalist industry. While agriculture generated both its own capitalists and workers, the urban crafts played a distinctly secondary role in forming either pole of industry. Rather, the agricultural revolution inadvertently supplied the labourers, and merchants advanced much of the money to employ them and shaped markets in which their products were sold. To grasp the birth of industrial capital, we must first look at merchants.

The question is how merchant activity fostered primitive accumulation. In the genesis of capitalism, Marx held that merchants played a decisive, independent role: 'Today, industrial supremacy brings with it commercial supremacy. In the period of manufacture it is the reverse: commercial supremacy produces industrial predominance' (1890, p. 918). Of course market growth need not stimulate either industry or wage-labour; it led to the development of grainproducing serfdom in Poland and slave sugar and tobacco plantations in much of the Americas. But even these might have contributed to capitalist development if trade with peripheral areas using these labour forms financed industrial production in England (Wallerstein 1976; cf. Brenner 1977).

Merchants could foster primitive accumulation by expanding markets, by providing employment, or by investing profits. While Marx emphasizes domestic causes of proletarianization, he focuses primarily on international commerce in accounting for the genesis of the industrial capitalist (1890, ch. 31). This interpretation stresses the forcefulness and unevenness of primitive accumulation; it was through servile labour in the colonies, the slave trade, and commercial wars that the English prospered and replaced the Dutch as the dominant mercantile power by 1700.

No doubt international commerce had a central role in industrial expansion. Growing exports stimulated domestic output; particularly for the textile industries, something like half the output of which was exported. In most of the eighteenth century, industrial exports grew more rapidly than industrial output, increasing their share of that output from about a fifth to a third from 1700 to 1800. Imports of industrial raw materials, like silk, cotton, dyestuffs and iron, also supported English industry. Marx's stress on the colonial system is warranted by the expansion of the share of domestic exports shipped to the American colonies from 11 per cent in 1700 to 37 per cent in 1772, as well as by its growing significance for imports and reexports (Davis 1962; Minchinton 1969; Cole 1981).

Merchant services and profits also stimulated domestic output. The ascendency of British merchants in world trade led to the expansion of the ports. Commerce was the principal factor in London's growth, and consumption spending by merchants, related professionals, and labourers fostered both industrial and agricultural expansion in much of England. Lesser ports had similar effects. Purchases of ships, armaments and connected products likewise supported industry. While large and growing, the reinvested profits of the international merchant community remained principally in the same lines of business and, except for a few industries in the ports, offered little industrial financing.

Marx's stress on international commerce is surely one-sided; others, including Lenin (1908), have shifted the focus to the home market. For this market, which in England regularly consumed some nine-tenths of the national product, grew with the perhaps 80 per cent increase of that product from 1700 to 1780.

But the home market had significance beyond its share of national output. As Hobsbawm argues (1954), capitalism involves production for a mass market, and the combination of traditional local and export markets could not supply the necessary scale. During the manufacturing period, an integrated, mass market was born. This transformation was not of course confined to the home market; Hobsbawm underscores the importance of new markets in the colonies. But the home market was primary. It became much more spatially integrated. For food, fuel and many industrial products, the great expansion of London was central to this process. Expanding national markets were accompanied by growing regional specialization of production. The mass market was supported by the emerging class structure, especially the prosperous middle class of farmers, modest merchants, manufacturers, and some professionals and tradesmen. Particularly in times of falling agricultural prices, workers added to this market. Finally, a series of new commodities spread through sections of the home market, including the new textiles, stockings, new tools, and a host of housewares made of metal, pottery and glass. For most of these, the home market was decisive (Eversley 1967: Thirsk 1978).

The reinvested profits of domestic merchants, like their international counterparts, remained preponderantly within the commercial sphere. They expanded their working capital, deepened their wholesale marketing network, and helped form the clearing-house and billdiscounting mechanisms through which the market worked. They were the principal investors in transportation improvements like the expansion of coastal shipping, turnpike construction, river deepening, and, from the second half of the eighteenth century, canal construction. Domestic merchants could also finance industry, but even if they did not, their investment created conditions where others would.

The Birth of Industrial Capital

In his well-known discussion of paths to capitalism, Marx identified two ways that industrial capitalists were formed; producers could become capitalists and merchants, or merchants could enter production and employ wage-labourers (1894, ch. 20: see also Dobb 1947). At stake is not just the genesis of industrial capital but also its dynamic. For Marx, the merchant path separates the worker from ownership of the product but retains inherited techniques and social organization of production. It is ultimately conservative; 'however frequently this occurs as a historical transition . . . it cannot bring about the overthrow of the old mode of production by itself, but rather preserves it and retains it as its own precondition' (1894, p. 452). By contrast, producers-turned-capitalists comprise 'the really revolutionary way' since they grow by transforming the organization and techniques of production.

Two quite different kinds of wholesale capitalist production were formed: manufacturing in the narrow sense and domestic industry. Manufacturing had the more innovative organization of the production process. It grouped craft workers specializing by task in the capitalist's workshop and often entailed economies of scale and significant capital costs. It was not solely the creation of producers; the funds, organizational abilities and market knowledge of merchants and even landlords also played a part. Manufacturing most commonly arose in industries which were new (alum, gunpowder, glass, cane sugar), used new techniques (salt, pig iron, heavy iron products), or produced for newly integrated markets (coal). Marx is ambiguous about its significance; he calls manufacture) 'a characteristic form of the capitalist process of production' which 'prevails throughout the manufacturing period' yet recognizes that it never dominated the system (1890, pp. 455, 911).

Domestic industry was far more widespread. Born earlier in the textile trades, domestic industry expanded across many industries in the manufacturing stage. Spurred by relatively high wages and inelastic labour supply in the organized urban trades, both merchants and producers put out work to be done in the homes of outworkers. Some domestic industry arose in urban areas, especially London, but more was proto-industrial – household production of wholesale industrial goods by those retaining ties to land and rural communities (Mantoux 1928; Mendels 1972; Kriedte et al. 1981).

This proto-industry had distinctive patterns of development. It generally originated in pastoral regions and declining or large-scale agricultural areas. Over time, outwork by independent producers declined and wage-labour rose. Through the efforts of both merchants and producers, proto-industry spread within and between localities. Immigration and a distinctive proto-industrial family structure which encouraged earlier marriages and rising birth rates gave an elasticity to employment in existing areas, but ties to the land meant that rapid expansion could only be achieved by the geographic spread of industry. Much of this growth was undertaken by the formation of new firms holding advantages of knowledge of and proximity to the local population.

By themselves or with others, producers were instrumental in changing the production process and its products. Both from the Continent and within England, craftsmen diffused techniques to make pig iron, paper, saltpetre and brass and copper products. They also made a few advances in coal mining, iron making and civil engineering. No doubt the division of labour was refined in manufacture and learning in the proto-industrial regions improved skills. But the circumscribed technical knowledge of most crafts and the personal interactions required to transmit skills formed barriers which limited the scope and importance of technical innovations and rendered their diffusion slow and uneven.

Changes in products were far more general. The largest of the rural industries, textiles, maintained its position in the world market by developing the new, lighter fabrics called the new draperies, as well as introducing cotton, fustian, linen and silk. Pots, pans, nails, pins, knives, buttons, stockings, ribbon, lace, glass bottles and earthen pots all developed for the home market in the late sixteenth and early seventeenth centuries (Thirsk 1978). Merchants and craftsmen were both active in developing and diffusing these product innovations.

These changes in techniques and products formed a dynamic in production which gave competitive advantages to innovating firms and regions. The use of these advantages helped replace the inherited pattern of local and external markets by a new kind of market, called by Polanyi the internal market (1944). The products of innovators substituted for imports and also extended the market absolutely, particularly among middle and lower class consumers. The extent of the internal market grew with market integration and the increased per capita income and consumption resulting from productivity increases and transportation improvements. As its share of national product grew, industry came to create more of its own demand. Success in the internal market provided the basis from which some commodities entered export markets. In the international economy, as well as in England, industrial advance was leading to commercial success.

Primitive Accumulation and the Stages of the Capitalist Economy

In England, more than in any country before or since, the manufacturing stage realized the preconditions of capitalist production. This stage also created conditions for the distinctive kind and pace of accumulation characteristic of the stage of large-scale industry. It thus satisfies the criterion that Gerschenkron employed to assess the usefulness of the concept of primitive accumulation: whether this prior accumulation aided the rapid growth associated with the onset of industrialization (1966, pp. 31–51).

Manufacturing did this not so much in the way Gerschenkron stresses, by the transfer of previously accumulated wealth to industrialists. The modest capital requirements of early factories and the primary role of producers in founding industrial firms – Marx's revolutionary path to capitalism – makes it difficult to justify the role of the prior accumulation of wealth in this way (Crouzet 1985). Far more important were marketing and transportation investments, which together with the agricultural revolution developed markets wide enough to warrant the extensive factory investment and the formation of a capital goods sector characteristic of the Industrial Revolution (Hobsbawm 1954). Moreover, the manufacturing period generated the proletariat to work in the factories and – through the development of milling techniques, new products like the clock, the printing press, firearms and the Newcomen engine, and the great expansion of the tool-making sector – supplied agents willing and able to solve the technological problems of industrialization.

But even in England, primitive accumulation was by no means identical to the processes of manufacturing stage. It involved processes prior to this stage, like the growth of towns and the elimination of serfdom. Nor was it completed within the manufacturing stage. For the persisting ties to the land, the structure of income distribution, and the inherited forms of labour limited the supply of labour-power, the extent of the market, and the growth of productivity (Levine 1975). It was left to the dynamic of the next stage to complete the) 'dissolution of the old economic relations of landed property', since 'only with the development of modern industry to a high degree does this dissolution at individual points acquire its totality and extent' (Marx 1973, p. 277).

Still, the extent to which the conditions of capitalist production were created within the stage of manufacturing made England unique. In it alone had the agricultural revolution taken 'the classical form' (Marx 1890, p. 876). The success of its industrialization reinforced its uniqueness by altering the process of primitive accumulation. Growing productivity and falling prices undercut the viability of proto-industrial and town craft producers at home and, through the growth of an export economy, abroad. The steamship and railroad overcame locational limits to competition. Separation from the means of production had become a consequence of the industrial stage of capitalism.

Moreover, for latecomers the prior generation of a supply of money capital and labour-power within their countries had less importance than in England. Primitive accumulation was internationalized; capital and labour-power both migrated more readily. New credit institutions and state policies could supply capital during the course of industrialization (Gerschenkron 1962). On the Continent, large-scale industry was often born while peasantries persisted. More extreme was the United States, which was already a major industrial power at the time its frontier closed, and which, in the absence of widespread separation of agricultural producers from the land, sold principally in the home market.

Capitalism is for Marx a world-historical system, not a set of autarkic national units. There can therefore be no stage of primitive accumulation which uniformly prepares the way for capitalism in each of these units. The very success of the kinds of processes which brought large-scale, industrial capitalism in England changed both the process of primitive accumulation elsewhere and the relation of these processes to capitalist expansion. By tying the concept of primitive accumulation to a periodization of capitalist development, Marx provides insight into both the classical case of the genesis of capitalism and the necessarily different forms this genesis took elsewhere.

See Also

- ► Feudalism
- ► Capitalism
- ► Mode of Production

Bibliography

Brenner, R. 1977. The origins of capitalist development: A critique of neo-Smithian Marxism. *New Left Review* 104: 25–92.

Chambers, J., and G. Mingay. 1966. *The agricultural revolution: 1750–1880*. London: Batsford.

Cole, W. 1981. Factors in demand 1700–80. In *The economic history of Britain since 1700*, ed. R. Floud and D. McCloskey. Cambridge: Cambridge University Press.

Crouzet, F. 1985. *The first industrialists: The problem of origins*. Cambridge: Cambridge University Press.

Davis, R. 1962. *The rise of the English shipping industry in the seventeenth and eighteenth centuries*. London: Macmillan.

Dobb, M. 1947. *Studies in the development of capitalism*. New York: International.

Eversley, D. 1967. The home market and economic growth in England, 1750–80. In *Land, labour and population in the industrial revolution*, ed. E. Jones and G. Mingay. London: Edward Arnold.

Gerschenkron, A. 1962. *Economic backwardness in historical perspective*. Cambridge, MA: Harvard University Press.

Hobsbawm, E. 1954. The general crisis of the European economy in the 17th century. *Past and Present* 54(4): 33–53; 54(6): 44–65.

John, A. 1965. Agricultural productivity and economic growth in England, 1700–1760. *Journal of Economic History* 25: 19–34.

Jones, E. 1974. *Agriculture and the industrial revolution*. New York: Wiley.

Kerridge, E. 1967. *The agricultural revolution*. London: George Allen & Unwin.

Kriedte, P., H. Medick, and J. Schlumbohm. 1981. *Industrialization before Industrialization: Rural industry in the genesis of capitalism*. Cambridge: Cambridge University Press.

Lenin, V.I. 1908. *The development of capitalism in Russia*, 2nd edn. In *Collected works*, vol. 3, ed. V.I. Lenin. Moscow: Progress, 1964.

Levine, D. 1975. The theory of growth of the capitalist economy. *Economic Development and Cultural Change* 24(1): 47–74.

Mantoux, P. 1928. *The industrial revolution in the eighteenth century*. New York: Macmillan, 1961.

Marx, K. 1890. *Capital: A critique of political economy*, vol. I. New York: Vintage, 1977.

Marx, K. 1894. *Capital: A critique of political economy*, vol. III. New York: Vintage, 1981.

Marx, K. 1973. *Grundrisse: Foundations of the critique of political economy*. Harmondsworth: Penguin.

Mendels, F. 1972. Proto-industrialization: The first phase of the industrialization process. *Journal of Economic History* 32(1): 241–261.

Minchinton, W. 1969. *The growth of English overseas trade in the seventeenth and eighteenth centuries*. London: Methuen.

Polanyi, K. 1944. *The great transformation: The political and economic origins of our time*. Boston: Beacon, 1957.

Smith, A. 1776. *An inquiry into the nature and causes of the wealth of nations*, ed. E. Cannan. New York: Random House, 1965.

Tawney, R. 1912. *The Agrarian problem in the sixteenth century*. New York: Harper, 1967.

Thirsk, J. 1978. *Economic policy and projects: The development of a consumer society in early modern England*. Oxford: Clarendon Press.

Tilly, C. 1984. Demographic origins of the European proletariat. In *Proletarianization and family life*, ed. D. Levine. New York: Academic Press.

Wallerstein, I. 1976. *The modern world system: Capitalist agriculture and the origins of the European world-economy in the sixteenth century*. London: Academic Press.

Principal and Agent (i)

J. E. C. Munro

Keywords

Asymmetric information; Principal and agent (i); Contract law; Munro, J. E.

JEL Classifications

D8

An agent is a person who is employed to do an act on behalf of another called the principal, so that as a rule the principal himself becomes bound. That one person can represent another is a doctrine that has developed but slowly. In Roman law it was a general principle that no one could enter into a contract by stipulation on behalf of another, and in the case of mandate the mandatarius or quasi-agent incurred a personal liability towards their parties. The modern principle is that contracts entered into by an agent are regarded as entered into by the principal, provided the contract is within the scope of the agent's authority.

No special form of words is required to appoint an agent, and agency may be inferred from the conduct of the parties. An agent is required to conduct the business entrusted to him with as much skill as is generally possessed by persons engaged in a similar business, to act with reasonable diligence, to display the utmost fidelity, to keep proper accounts, and to pay over all moneys received less any expenses and his own remuneration.

Directors, managers, clerks, and servants, having power to act for their principals or masters, are agents. Besides these, the chief classes of agents are (*a*) factors; (*b*) brokers; (*c*) auctioneers ; and (*d*) ship masters. Each class is subject to the usages of the trade relating to the class. An agent cannot as a rule delegate his powers, but by the custom of certain trades sub-agents may be employed. The relation of principal and agent is terminated by mutual consent, by revocation, by the agent renouncing, by the expiration of the time agreed upon by the completion of the business, by the death or lunacy of either principal or agent, and by the bankruptcy of the principal.

Principal and Agent (ii)

Joseph E. Stiglitz

Abstract

The principal–agent literature is concerned with how the principal (say an employer) can design a compensation system (a contract) which motivates another individual, his agent (say the employee), to act in the principal's interests. A principal–agent problem arises when there is imperfect information concerning what action the agent either has undertaken or should undertake. It arises in insurance and credit relationships because of their intertemporal nature, when it is known as 'moral hazard'. It also arises where opportunities exist for the principal to extract as much rent as possible from the agent.

Keywords

Asymmetric information; Compensation; Competitive equilibrium; Credit; Employer–employee relationship; Implicit contracts; Individual rationality constraints; Insurance; Intertemporal relationships; Mirrlees, J.; Moral hazard; Optimal taxation structures; Pareto efficient tax structures; Partially discriminating monopoly; Pooling equilibrium; Principal and agent (ii); Rent seeking; Reservation utility constraint; Risk; Risk sharing; Ross, S; Separating equilibrium; Sharecropping; Stiglitz, J. E

JEL Classifications

D0

The principal–agent literature is concerned with how one individual, the principal (say an employer), can design a compensation system (a contract) which motivates another individual, his agent (say the employee), to act in the principal's interests. The term principal–agent problem is due to Ross (1973). Other early contributions to this literature include Mirrlees (1974, 1976) and Stiglitz (1974, 1975).

A principal–agent problem arises when there is imperfect information, either concerning what action the agent has undertaken or should undertake. In many situations, the actions of an individual are not easily observable. It would be very difficult for a landlord to monitor perfectly the weeding activity of his tenant. A bank cannot monitor perfectly the actions of those to whom it lends money. The employer cannot travel on the road with his salesman, to monitor precisely the effort he puts into his salesmanship. In each of these situations, the agent's (tenant's, borrower's, employee's) action affects the principal (landlord, lender, employer). Clearly, if an individual's actions are unobservable, then compensation cannot be based on those actions. In some cases, even if an individual's actions are not directly observable, it may be possible to infer his actions. Thus, if output were a function just of effort [$Q = F(e)$] then even if effort were unobservable, if output were observable, and the relationship between output and effort were known, then effort could be inferred with perfect accuracy.

The principal–agent literature focuses on situations where an individual's actions can neither be observed nor be perfectly inferred on the basis of observable variables; thus, for instance, it is usually assumed that output is a function of effort and an unobservable random variable, $\theta : Q = F(e, \theta)$.

Moreover, in many circumstances, the principal wishes the agent to take actions based on information which is available to the agent, not the principal. Indeed, this is the very reason that individuals delegate responsibility. Because of the asymmetry of information, the principal does not know whether the agent undertook the action the principal would himself have undertaken, in the given circumstances. Hence, even if the principal can observe the action, he may not know whether that action was appropriate.

Since, in general, the pay-offs to the agent will differ from those to the principal, the agent will not in general take the action which the principal would like him to take, or that they would contract for in the presence of perfect information. For instance, the employee may not adjust his effort as the situation requires, or he may engage in too much or too little risk taking.

The principal–agent problem is, then, the central problem of economic incentives.

In spite of the importance attached to *economic incentives*, until recently economic theory had little to say on the matter. In the standard theory, individuals were paid for performing a particular task. If they performed the task, they received their compensation; if they failed to perform the task, they did not. Individuals thus always had an incentive to perform the contracted-for service. Only if the employer were so foolish as to pay the worker whether he performed the task or not would an incentive problem arise.

The standard theory was based on the assumption that what action the 'principal' wished his agent to perform was perfectly known, and that the action could be perfectly and costlessly monitored. Neither assumption is plausible and, indeed, relatively few workers are paid solely on the basis of their observed inputs.

Origins of Principal–Agent Problems

Principal–agent problems arise whenever one individual's actions have an effect on another individual. The question arises, then, why cannot economic relationships be designed to avoid this kind of dependency? Under what circumstances do these interdependencies arise? For instance, if a landlord were to sell or rent his land to his tenant, then the workers' effort would have no effect on him. If an employer were to sell or rent his capital to his worker, then the workers' effort would again have no effect on him. Traditional neoclassical analysis emphasized the symmetry in economic relationships: one could describe the employer–employee relationship as the employee hiring capital just as well as one could describe it as the employer hiring labour. (This Wicksellian description of economic relationships always seemed peculiar to me; it seemed to suggest the absence within neoclassical analysis of certain important aspects of economic relationships; it is those aspects which are the subject of scrutiny here.)

There are three important reasons for the existence of principal–agent problems. Two have to do with the essential intertemporal nature of certain relationships: insurance and credit. When two individuals enter into an insurance contract, one individual *(a)* promises to pay the other (*b*) a certain amount if event A occurs, while the other (*b*) promises to pay *(a)* a certain amount if event B occurs. If there are actions which one of the individuals can undertake between the date of the contract and the event which will affect the outcome, then there is a principal–agent relationship between the two. This particular form of the principal–agent problem is referred to within the insurance literature as the *moral hazard problem* (see Arrow 1965), and, by extension, the term has been applied to the principal–agent problem more generally.

Similarly, in credit relationships, one individual gives another some resource (money), in return for a promise to repay that money at some later date. So long as there is some probability of default, which can be affected by the actions of the borrower, there is a *moral hazard* or *principal–agent* problem (provided that that action cannot be perfectly monitored by the lender).

Many economic relationships have an important element of insurance within them. The landlord–tenant sharecropping relationship can be viewed as if the tenant pays a fixed rent, and then receives an insurance policy from the landlord, in which the landlord agrees to pay the tenant a certain amount if output is low (equal to the difference between his share and the fixed rent); and the tenant agrees to pay a premium equal again to the difference between the share and the fixed rent, when output is high.

Indeed, the credit 'problem' can be viewed as a special form of an insurance relationship: the lender provides an insurance policy, such that if the borrower's resources are less than the amount owed, the lender agrees to pay the borrower the difference (which the borrower then immediately repays to the lender). The premium is the difference between the rate of interest on a perfectly safe loan and the rate of interest charged on this risky loan.

Insurance (spreading and transferring risk) provides one of the explanations of sharecropping; were workers to rent the land, they would have to absorb all the risk associated with output variations. With sharecropping, the risk is shared between the landlord and the tenant. Since the wealth of tenants is usually much less than that of landlords, there is some presumption that the landlords are better able to absorb this risk.

But even if the tenants were risk neutral, there might be a principal–agent problem. We suggested above that if the landlord were to rent his land to the tenant, there would be no principal–agent problem. But this is not quite correct. If the tenant did not have sufficient resources to pay the rent before production, then the landlord would have to lend the tenant the money. (If he receives the rent at the end of the period, then it is as if he is lending the individual the money). And then, if there are actions which the individual can undertake which affect the likelihood of not being able to repay the debt (pay the rent), *then* there is a moral hazard problem.

There is a second reason that renting land might not solve the moral hazard problem. There may be actions which the tenant can take which affect the quality of the land. To the extent that those actions are monitorable, the rental agreement may specify the actions to be undertaken (e.g. concerning what crops are to be grown, or grazing patterns). But these actions are not perfectly monitorable, and thus, even with rental agreements there are important principal–agent problems. (The same issues arise, of course, with the rental of any durable good).

Again, one should ask, cannot these principal–agent problems be alleviated, e.g. by *selling* the asset. But this entails precisely the two problems we identified before as giving rise to principal–agent relationships: The agent (tenant, employee) may not have sufficient capital (and thus must borrow to make the purchase, creating a credit principal–agent problem); and if there is any risk associated with the future value of the land, it imposes a risk on the agent. Any attempt to alleviate those risks (through insurance) again gives rise to a moral hazard problem.

The third major source of principal–agent relationships is rather different. It arises from the attempt of the principal to extract as much *rent* (surplus) from the agent as possible. The employer does not know how difficult the task is that he would like the worker to perform. He could pay the worker the full value of his output, but that would leave him no profits. He might pay much less, but that might result in the worker refusing to work, if the task is in fact quite difficult; and thus he would lose profits that he might otherwise obtain. This rent extraction problem has been particularly well studied in the context of public utilities: the government does not know the minimum amount of compensation required to keep the utility producing. The rent extraction problem may be alleviated within competitive environments by holding auctions: the individual for whom the asset (franchise) is most valuable will bid the most. But there may not be enough bidders to extract all the rents through an auction mechanism; and at least in the case of utilities, the government may care not only about the rents received, but also about the actions undertaken by the franchisee. (In some cases, the rent extraction problem and the insurance problem are closely related: the average value of rents received may be increased if rents can be varied with the weather, the state of nature; again, we can think of decomposing the rent payment into a fixed rent and an insurance payment).

This list of reasons for the origins of principal–agent relations is not meant to be exhaustive; yet many of the other reasons cited may be reduced to one of these explanations. For instance, consider the problem of a production line on which there are many workers; the output of the production line depends on all of their efforts. In the absence of risk aversion and credit problems, the incentive problem could be solved by giving each worker the total value of net output. He would purchase the right to the job by paying a fixed fee. With such a compensation scheme, the worker would have full incentives for maximizing the firm's output. But such a compensation scheme imposes on the worker an intolerable level of risk; and the fixed fee he would be required to pay necessitates his borrowing large amounts of money.

The Basic Principal–Agent Problem

In the standard principal–agent problem, one looks for that contract (compensation scheme) which maximizes the expected utility of the principal, given that (a) the agent will undertake the action(s) which maximizes his expected utility, given the compensation scheme; and (b) given that he must be willing to accept the contract.

The second set of constraints (which are nothing more than the standard reservation utility constraints) are sometimes referred to as the individual rationality constraints.

There are two standard mathematical formulations. One is a direct generalization of the insurance–moral hazard problem. There are a set of observable events, such as whether an accident occurs. The probability that an event *i* occurs is a function of the actions undertaken (effort at accident avoidance):

$$p_i = p_i(\mathbf{e}),$$

where **e** may be a vector. The wealth of the individual in state *i*, in the absence of insurance, is w_i, and with insurance it is y_i. Thus

$$h_i = y_i - w_i$$

is the net payment from (to) the insurance company (the principal) in state *i*. The expected utility of the insured (the agent) is then just

$$U = \sum_i U_i(y_i, \mathbf{e}) p_i(\mathbf{e})$$

while that of the principal is

$$V = \sum_i V_i(h_i) p_i(\mathbf{e}).$$

$\{\mathrm{h_i}\}$ is chosen to maximize V subject to $U \geq \overline{U}$.

Notice that the employer–employee relationship may be cast in this form: the observable events are the levels of output. Assume for simplicity, that we measure outputs in round numbers (say, bushels of wheat). Then state *i* refers to the number of bushels produced. p_i then is the probability that *i* bushels will be produced. Assume that the individual's wealth, apart from this contractual arrangement with his employer, is zero. Then y_{i} is the individual's pay if output is *i*. If the employer is risk neutral,

$$V_{\mathrm{i}}(h_i) = qi - hi = qi - y_i,$$

where q is the price of output (of a bushel of wheat), assumed to be independent of *i*.

Although the employer–employee relationship can be cast in this form, it is more naturally represented by a formulation in which the probabilities of the states (weather) are fixed, where the states are unobservable, but where what is affected by the employee is the output in each state.

We can represent this formally in the following way. Let *S* be a set of state variables (like weather) observable to the agent. Let *Q* be a set of output variables (assumed observable to the principal and agent). And let *A* be a set of inputs (actions) by the agent assumed observable only by the agent.

Then a compensation scheme is a payment from the principal to the agent which is a function of all variables that are observable to both the agent and principal.

$$Y = \varphi(Q)$$

The agent chooses his actions to maximize his expected utility which depends both on his income and his actions, given

$$\max \; EU(Y, A, S)$$

where outputs (actions), *A*, are related to the inputs by a production function

$$Q = Q(A, S)$$

We denote the solution to this by

$$A = H(S).$$

Finally, we can calculate the expected utility of the principal; his utility depends on the agent's actions, the payments he makes to the agent, and his state (the actions may affect the principal either directly, or via their effect on outputs, or via their effects on payments).

$$EV = EV(\varphi(Q), Q, A, S).$$

The principal's problem is to choose φ to maximize his expected utility

$$\max \; EV$$

recognizing the dependence of the agent's action on φ and recognizing that he must pay the agent enough to induce him to accept the job

$$EU \geq \overline{U} \qquad \text{(RU)}$$

Pooling Versus Separating Equilibrium

Much of the literature has focused on situations where the principal wishes to induce the agent to take different actions in different states. That is, in the simplest case where only output is observable by the principal, if $A^*(S)$ is the action desired in state S, then the compensation scheme must be such that

$$EU[\varphi(Q(A^*, S)), A^*, S] > EU[(Q\text{-}(A, S)), A, S] \quad \text{for all feasible } A.$$

These constraints are referred to as the self-selection or incentive compatibility constraints.

When the individual takes actions in two different states, so that the observable variables are the same, i.e. so that the principal cannot distinguish which of the two states has occurred, we say that there is a *pooling* equilibrium. When the individual takes actions so that the principal can identify which state has occurred, we say that there is a *separating* equilibrium. (This terminology was introduced within the context of the adverse selection literature by Rothschild and Stiglitz (1976)). A basic result of the principal–agent literature establishes conditions under which the optimal contract involves complete or partial separation.

Adverse Selection

The variable S can be thought of as a characteristic of an individual, rather than as the state of nature. Then the self-selection constraint says that individuals of type S prefer action $A(S)$ to any other feasible action. If the self-selection constraints are satisfied, we can identify who is of what type. The action may consist of nothing more than making a choice. In the adverse selection interpretations of the model, the constraint (RU) needs to be replaced by the set of constraints,

$$U(\varphi(Q(A,S)),A,S) \geq \overline{U}(S), \quad \text{for all } S$$

that is, there is a reservation utility level for each individual (an individual rationality constraint for each type). (Note that a similar set of constraints is relevant if the contractual arrangement between the principal and agent is not binding, i.e. the individual can quit after he sees what the state of nature is).

Some examples follow.

i. *The partially discriminating monopoly* (see, e.g. Salop 1977; Stiglitz 1977). The firm knows that different individuals have different indifference curves between the good he sells and other goods, and different reservation utility levels, but he does not know who is of which type. Q may be the quantity of some commodity chosen by an individual, in which case φ (Q) can be interpreted as the payment to the monopolist. (If one individual unambiguously has stronger preferences for the good, in the sense that at any quantity and payment, the extra amount he is willing to pay for a marginal unit is greater, then some separation is always desirable; this property is called the single crossing property).
ii. *Optimal tax structures* (Mirrlees 1971). The government wishes to impose differential taxation on different individuals; it may want to impose a higher tax on the more able, but cannot tell who is the more able. Neither the individual's productivity nor the number of hours a week he works is observable, but his income is observable. The income tax schedule specifies a level of consumption corresponding to each level of income. The individual chooses (by the amount of work he undertakes) a point on that schedule. A schedule which results in the more able earning (choosing) higher incomes is one which separates. This will be desirable if the indifference curves between consumption and income are flatter for the more able – they require less of an increase in consumption to compensate for an increase in income. This will be true, for instance, if the underlying indifference curves between hours worked and consumption are the same for all individuals.
iii. *Pareto efficient tax structures* (Stiglitz 1982a). In the previous problem, the government maximized the sum of utilities, subject to the self-selection constraints, the revenue constraints, and the individual rationality constraints (which simply required that the individual desire to work). The revenue constraint was equivalent, in this problem, to the profits

(revenues) of the landlord; that is, while in the landlord problem we maximize the revenue, subject to the expected utility of the individual satisfying a certain constraint, here the dual of this problem is analysed. The 'sum of utilities' is equivalent to 'expected utility' – where the probability of each state S is identical. We can directly generalize this by imposing constraints on the level of utility attained by all individuals other than the first; we then maximize the first individual's utility subject to these constraints (and subject to the self-selection constraints, and the revenue constraints). This is the problem of Pareto efficient taxation. It is equivalent to the problem of maximizing a weighted sum of individuals' utilities.

iv. *Implicit contracts with asymmetric information*. (For surveys, see Hart 1983; Stiglitz 1986; Azariadis and Stiglitz 1983). With perfect information, the employer would provide insurance to the employee, to stabilize the employee's income. If, for instance, the workers' utility function was separable between hours worked, l, and income y

$$U = u(y) - v(l),$$

then with complete information, and risk neutral firms, y will be the same in all states, but l will be higher in states where labour productivity is higher. Thus, if the employer knew the state, but the worker did not, the employer would have an incentive always to say that it was a good state (since what he paid the worker was the same, but workers are required to work more in good states). The optimal contract will induce the employer to announce that it is bad when it is in fact bad, i.e. it will separate (at least partially).

Qualitative Results

It is clear that many economic relationships fall within the scope of the 'principal–agent' model. Many of the basic qualitative results emerge from a detailed analysis of the insurance model:

(a) There is a risk-incentive trade-off; since the risks undertaken will be a function of the quantity of insurance purchased, if the latter is observable, the premium will depend on it, and in equilibrium, there will be quantity rationing, i.e., the individual would like to purchase more insurance, at the going benefit premium ratio (Pauly 1968). The amount of insurance will be greater, the more risk averse the individual.

(b) Indifference curves (between benefits and premia) are not generally quasi-concave, nor feasibility sets (the set of insurance premia satisfying the nonnegative profit constraint) convex; this has important consequences for the existence of competitive equilibria. The amount of insurance purchased may not be a continuous function of the price of insurance; and the level of effort may not be a continuous function of the amount of insurance purchased.

(c) Competitive equilibrium, when it exists, will not in general be Pareto efficient (Arnott and Stiglitz 1986; Greenwald and Stiglitz 1986); the profits of one insurance firm are affected both by the terms at which other firms offer insurance contracts (whether for similar accidents or not), and by the prices at which goods (whether complements or substitutes for accident avoidance or accident inducing activities) are sold; there exist a set of Pareto improving subsidies and taxes. In some instances, firms may attempt to internalize some of these 'externalities.' This leads to interlinkage of markets, both across time (the same insurance firm insures the individual over time), and at the same time (the same insurance firm insures the individual for many different risks) (Braverman and Stiglitz 1982). The frequently observed interlinkage between credit and land markets in less developed countries has been interpreted in this light.

Variants of the General Model

Further results have been obtained for various variants of the general model. We discuss a few of the more important versions below:

i. Adverse selection model. The major qualitative results of this model (other than the specification of the conditions under which pooling or separation occurs, discussed above) entail an analysis of the distortions (relative to perfect information) engendered by the self-selection constraints; in the optimal income tax, the reduction in work (income) of the less able (associated with a positive marginal tax rate); in the asymmetric information implicit contract model, in the existence of overemployment in good states (with the separable utility function and risk neutral firms), or underemployment in bad states (with very risk averse firms). To discriminate among individuals, firms may engage in socially wasteful activities, such as random pricing or long queues. Generally, one group in the population (the most risk averse in the insurance model, the highest ability in the optimal income tax model) chooses a contract which does not distort its behaviour.

ii. Incentive model with actions taken before state is known. When the random elements have bounded support, then a first best can be achieved simply by imposing a large enough penalty for performances below a given threshold. The individual will exert enough effort to avoid this. (See Mirrlees 1974; Stiglitz 1975).

iii. Theory of contests. If the output of others performing similar tasks in similar situations is observable, then one will employ compensation schemes based on relative performance; these will do better than individualistic compensation schemes. If there are enough individuals, simple schemes, based only on individuals' rankings, can approximate the first-best outcomes.

iv. Models in which the utility constraint is not binding. In some cases, when the principal maximizes his expected utility, subject to the workers' reservation utility constraint, the latter constraint will not be binding. Such models give rise to unemployment. A particularly important variant of these models is described next.

v. Models in which quality is affected by price. If the probability of default increases with the rate of interest charged (either because individuals undertake more risks when the interest rate is higher, or because those who are less risky stop applying for loans at high interest rates), then banks may not raise interest rates, even in the presence of an excess demand for loans. Similarly, if the productivity of a worker increases with the wage paid (either because individuals exert greater effort at higher wages or because those who are recruited at higher wages are more productive), then firms may not lower wages, even in the presence of an excess supply of labour.

vi. Terminations. In multiperiod models, it has been shown that the optimal contract may entail the termination of a relationship when performance is unsatisfactory; this is shown to be preferable to the imposition of other penalties. (See Stiglitz and Weiss 1983).

vii. Infinite period models. Long-term relationships may ameliorate some of the incentive problems (see Radner 1981). Over an infinite lifetime, the principal (insurer) can make good inferences concerning the actions of the agent (insured); the relative frequency of accidents will converge to the accident probability corresponding to the individual's effort level. Not surprisingly, then, with low enough discount rates, incentive schemes can be designed which approximate the first best outcomes. The interpretation of this result is, however, subject to some controversy. Since with low discount rates, the change in lifetime income which would be associated with the individual bearing the full risk of the outcome for any period is negligible, it is as if the individual is risk neutral; and with risk neutrality we know that first best optimum can be obtained (if bankruptcy is ignored).

The Set of Admissible Contracts

One of the important and general results to emerge from the principal–agent literature is that the nature of the equilibrium contract depends on the set of admissible contracts. Contracts can depend only on the available information; typically, it is desirable to use all of the available information, though in practice, many variables which ought to be relevant (have information value) are not included within the compensation scheme.

Similarly, if one could costlessly implement a non-linear incentive scheme, such schemes would, in general, be preferable to linear schemes. Though in practice, again, most observed schemes seem relatively simple (linear, piece-wise linear, etc.), much of the literature has been concerned with characterizing in admittedly simple situations the optimal non-linear scheme.

In a variety of situations, if one could make pay a stochastic function, it would be desirable to do so, even with risk averse individuals. (Arnott and Stiglitz (1988) and Holmstrom (1979) show that with separable utility functions, this will not happen). The intuition behind this, in the case when actions have to be taken prior to the agent obtaining information about the state is that the possibility that he receives a low compensation so induces him to work hard that the employer (landlord) can reduce the dependence (on average) of pay on output, and thus reduce the variability of income.

Though optimal schemes may thus appear to be fairly complex, in practice most schemes employed are relatively simple. There is an ongoing controversy between those who seek to consider increasingly complex schedules, dismissing work which has analysed simple linear schedules as ad hoc; and those who seek to explain the kinds of compensation schedules actually employed; these dismiss the complex solutions as being irrelevant. They would argue that efforts should be devoted to understanding why actually employed schemes take on the form they do.

One possible explanation of the use of simple schedules is that they may be more *robust*. That is, as technology changes or the probability distribution of states changes (the exogenous parameters in the principal–agent problem) the optimal compensation scheme changes. But in practice, revisions to compensation schemes are costly, and one must find a scheme that works under a variety of situations. Simple, linear schemes may possess this property of robustness.

Another important characteristic of the set of admissible schemes relates to commitments. Can, for instance, the worker commit himself not to leave, or can the employer commit himself not to terminate the relationship?

A closely related issue is the set of punishments (rewards) which are admissible. It makes a great deal of difference if there are limits on the negative compensations that can be provided in the presence of bad outcomes.

We have noted the role of observability in the design of contracts. In some cases an important distinction may arise between observability and verifiability. The question is associated with how a contract is to be enforced. If the contract is to be enforced through the courts, it must be the case that any violation can be verified by an outside third party. Both the principal and the agent might know that the contract has been violated i.e. they both may observe the S (and not S') has occurred, and therefore that the payment should be that corresponding to S (and not S'). But unless it can be proved, the principal might attempt to cheat the agent. Knowing this, the agent would refuse to sign a contract based on unverifiable variables.

On the other hand, if the contract is enforced by a reputation mechanism, good behaviour may be enforced so long as the state is observable by both parties.

Concluding Remark

We have focused here on a discussion of general principles. It should be emphasized, however, that the principal–agent model has provided important insights into the nature of a variety of economic relationships, in labour, land, credit, and product markets. These detailed applications of the general theory represent an important area of on-going research.

See Also

▶ Incomplete Contracts

Bibliography

Allen, F. 1981. The prevention of default. *Journal of Finance* 36: 271–276.

Arnott, R., and J.E. Stiglitz. 1982. The welfare economics of moral hazard. Discussion Paper No. 465, Queen's University, Kingston, Ontario, March.

Arnott, R., and J.E. Stiglitz. 1985. Labor turnover, wage structures, and moral hazard: The inefficiency of competitive markets. *Journal of Labor Economics* 3: 434–462.

Arnott, R., and J.E. Stiglitz. 1986. Moral hazard and optimal commodity taxation. *Journal of Public Economics* 29 (1): 1–24.

Arnott, R., and J.E. Stiglitz. 1988. Randomization with asymmetric information: A simplified exposition. *RAND Journal of Economics* 19: 344–362.

Arrow, K.J. 1965. Aspects of the theory of risk-bearing. Helsinki: Yrjö Jahnsson Foundation.

Azariadis, C., and J.E. Stiglitz. 1983. Implicit contracts and fixed price equilibria. *Quarterly Journal of Economics* 98 (3.) Supplement: 1–22.

Braverman, A., and J.E. Stiglitz. 1982. Sharecropping and the interlinking of agrarian markets. *American Economic Review* 72: 695–715.

Eaton, J., and M. Gersovitz. 1981. Debt with potential repudiation; theoretical and empirical analysis. *Review of Economic Studies* 48: 289–309.

Fellingham, J.C., Y.K. Kwon, and D.P. Newman. 1982. Ex ante randomization in agency models. *Rand Journal of Economic Studies* 15: 290–301.

Gjesdal, F. 1982. Information and incentives: The agency information problem. *Review of Economic Studies* 49: 373–390.

Green, J., and N. Stokey. 1983. A comparison of tournaments and contests. *Journal of Political Economy* 91: 349–364.

Greenwald, B., and J.E. Stiglitz. 1986. Externalities in economies with imperfect information and incomplete markets. *Quarterly Journal of Economics* 101 (4): 229–264.

Grossman, S., and O. Hart. 1983. An analysis of the principal–agent problem. *Econometrica* 51 (1): 7–45.

Hart, O. 1983. Optimal labor contracts under asymmetric information: An introduction. *Review of Economic Studies* 50: 3–35.

Helpman, E., and J.J. Laffont. 1975. On moral hazard in general equilibrium. *Journal of Economic Theory* 10: 8–23.

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

Holmstrom, B. 1982. Moral hazard in teams. *Bell Journal of Economics* 13: 324–340.

Keeton, W.R. 1979. *Equilibrium credit rationing*. New York: Garland Publishing.

Lazear, E., and S. Rosen. 1981. Rank order tournaments as optimum labor contracts. *Journal of Political Economy* 89: 841–864.

Mirrlees, J. 1971. An exploration of the theory of optimum income taxation. *Review of Economic Studies* 38 (2): 175–208.

Mirrlees, J. 1974. Notes on welfare economics, information, and uncertainty. In *Contributions to economic analysis*, ed. M.S. Balch, D.L. McFadden, and S.Y. Wu. Amsterdam: North-Holland.

Mirrlees, J. 1976. The optimal structure of incentives and authority within an organization. *Bell Journal of Economics* 7 (1): 105–131.

Nalebuff, B., and J.E. Stiglitz. 1983a. Information, competition and markets. *American Economic Review* 72: 278–284.

Nalebuff, B., and J.E. Stiglitz. 1983b. Prizes and incentives: Towards a general theory of compensation and competition. *Bell Journal of Economics* 14: 21–43.

Pauly, M.V. 1968. The economics of moral hazard: Comment. *American Economic Review* 58: 531–536.

Radner, R. 1981. Monitoring cooperative agreements in a repeated principal–agent relationship. *Econometrica* 49: 1127–1148.

Radner, R., and J.E. Stiglitz. 1983. A nonconcavity in the value of innovation. In *Bayesian models in economic theory*, ed. M. Boyer and R. Kihlstrom. Amsterdam: North-Holland.

Ross, S. 1973. The economic theory of agency: The principal's problem. *American Economic Review* 63: 134–139.

Rothschild, M., and J.E. Stiglitz. 1976. Equilibrium in competitive insurance markets: An essay on the economics of imperfect information. *Quarterly Journal of Economics* 90: 629–649.

Salop, S. 1977. The noisy monopolist: Imperfect information, price dispersion and price discrimination. *Review of Economic Studies* 44: 393–406.

Salop, S., and J. Salop. 1976. Self-selection and turnover in the labor market. *Quarterly Journal of Economics* 90: 619–628.

Sappington, D., and J.E. Stiglitz. 1987. Information and regulation. In *Public regulation: New perspectives on institutions and policies*, ed. E. Bailey. Cambridge, MA: MIT Press.

Spence, A.M., and R. Zeckhauser. 1971. Insurance, information, and individual action. *American Economic Review* 61: 380–387.

Stiglitz, J.E. 1974. Incentives and risk sharing in sharecropping. *Review of Economic Studies* 41: 219–255.

Stiglitz, J.E. 1975. Incentives, risk, and information: Notes toward a theory of hierarchy. *Bell Journal of Economics* 6: 552–579.

Stiglitz, J.E. 1977. Monopoly non-linear pricing and imperfect information: The insurance market. *Review of Economic Studies* 44: 407–430.

Stiglitz, J.E. 1982a. Self-selection and Pareto efficient taxation. *Journal of Public Economics* 17: 213–240.

Stiglitz, J.E. 1982b. Utilitarianism and horizontal equity: The case for random taxation. *Journal of Public Economics* 18: 1–33.

Stiglitz, J.E. 1986. Theories of wage rigidity. In *Keynes' economic legacy*, ed. J. Butkiewicz, K. Koford, and J. Miller, 153–221. New York: Praeger Publishers.

Stiglitz, J.E. 1987. On the causes and consequences of the dependence of quality on price. *Journal of Economic Literature*: 271–248.

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

Stiglitz, J.E., and A. Weiss. 1983. Incentive effects of terminations: Applications to the credit and labor markets. *American Economic Review* 73: 912–927.

Weiss, L. 1976. On the desirability of cheating, incentives and randomness in the optimal income tax. *Journal of Political Economy* 84: 1343–1352.

Principal Components

T. Kloek

The principal components of a set of m variables are m artificially constructed variables with the following properties. The first component 'explains' as much as possible of the total variance of the original variables. The second has the same property under the additional condition that it is uncorrelated with the first, and so on. It often happens that a few principal components account for a large part of the total variance of the original variables. In such a case one may omit the remaining components. The effect is a substantial reduction of the dimension of the problem. The method is used to explore the relations present in a set of data or to combat the problems created by multicollinearity.

As in linear regression, several approaches are possible. One may view the principal components as the solution to a simple mathematical plane fitting problem, or one may assume a statistical model with an unknown covariance matrix, which is to be estimated. A normality assumption may (but need not) be added, with the consequence that the method of maximum likelihood is available.

If we have a statistical model with an m-vector of random variables ε with covariance matrix Σ the kth principal component can be defined as $\pi_k = \varepsilon' a_k$ where a_k is the eigenvector (characteristic vector) of Σ that corresponds to the kth eigenvalue (characteristic root, latent root), the eigenvalues λ_k being arranged in descending order

$$\lambda_1 \geqslant \lambda_2 \ldots \geqslant \lambda_m. \tag{1}$$

If Σ is estimated by S the same operations of taking eigenvectors and eigenvalues are carried out with respect to S. The mathematics of this approach is explained in almost every book on multivariate statistical analysis. A classic in this field is Anderson (1958). In the econometrics literature a detailed account is given in Dhrymes (1970).

A descriptive approach starts with an $n \times m$ matrix X consisting of n observations on each of m variables. Then the principal components are the columns of $P = XA$, where A is the matrix of eigenvectors of $X' X$. As in (1) the eigenvectors are always arranged according to the descending order of the eigenvalues. The first principal component $p1$ may also be obtained as the result of minimizing the sum of the squares of the residuals E defined as

$$E = X - pa' \tag{2}$$

where p is an n-vector and a an m-vector. This approach to the subject is described in detail by Theil (1971, 1983).

Since both p and a are unknown we need an additional constraint in order to obtain unique results. Most authors choose $a' \ a = 1$ some $p' \ p = 1$. The choice is arbitrary and a matter of convenience. Here, it is henceforth assumed that $A' \ A = I$ and more generally that $A' \ A = I$. Another consequence of the fact that both p and a are unknown is that our problem does not have the simple linear structure of least squares regression. Hence the resulting A and P depend (in a non-trivial way) on the origin and scale of the original variables. In the statistical approach the variables are usually measured from their means, in the descriptive approach this is not always the case. If all variables are measured in the same units, there is a natural solution of the problem of the units of measurement. If this is not the case one often chooses the solution to take correlations rather than covariances. (This holds for Σ and S in the statistical approach but it may also be applied to $X' X$ in the descriptive approach.)

Geometrically, the principal components transformation is equivalent to rotating the scatter (in the descriptive approach) or the density (in the statistical approach). Consider the case $m = 2$ and suppose that the scatter has the form of an ellipse. Then the principal components transformation is equivalent to rotating the ellipse in such a way that the principal axes of the ellipse coincide with the axes of the coordinate system. Equivalently, one might rotate the coordinate axes in such a way that they coincide with the principal axes of the ellipse. More details on the geometry of principal components are given by Fomby et al. (1984, pp. 287–93).

The main purpose of applying principal components is *reduction of the dimension* of a data set. The idea originated with Hotelling (1933) and in the present author's opinion it can be interpreted as a mathematician's reaction on Thurstone's (1931) paper on factor analysis. Indeed, Hotelling applies his approach to psychological test scores. It was precisely for this type of data the psychologists developed factor analysis.

The main difference between factor analysis and principal components can be given as follows. In factor analysis it is assumed that Σ can be decomposed as:

$$\Sigma = CC' + D$$

where C is an $m \times h$ matrix and D a diagonal matrix of order $m \times m$. If

$$h < \frac{1}{2}\left[2m + 1 - \surd(8\mathrm{m} + 1)\right]$$

this assumption implies restrictions on the elements of Σ, while the principal components approach does not impose any restrictions on Σ.

A well-known economic example of dimension reduction was given by Stone (1947), who took 17 time-series from the US national accounts in the period 1922–38. They describe several income and expenditure aggregates relating to consumers, producers and the government. It appeared that in this period the first three principal components accounted for more than 97 per cent of the total variance of these 17 series (the first 80.8 per cent, the second 10.6 per cent, the third 6.1 per cent). The first principal component appears to be highly correlated with total income, the second with the annual change in income, the third with time. It should be emphasized that usually such simple interpretations are not available. More details are given by Stone (1947); also by Theil (1971).

Dimension reduction may also be desirable in the so-called *undersized sample* problem. Consider a (linear) simultaneous equation model. Suppose one wants to estimate the parameters of a simple structural equation by means of two-stage least-squares or a similar method. Then the first step requires the regressions of the current endogenous variables at the right hand side of the equation on the total set of predetermined variables. This is impossible if $n < m$ (the number of predetermined variables), but it may also have undesirable properties if $n < 2.5\ m$, say. In large models, but even in models of medium size, these rules may be violated. Kloek and Mennes (1960) proposed to tackle this problem by replacing the m predetermined variables by a limited number of principal components. For a further discussion and modifications, see Amemiya (1966). The limitations of this approach were discussed by Fisher (1965).

Dimension reduction may also be desirable in more general regressions with *multicollinear* explanatory variables. The principal components of these variables can play a very useful role in clarifying the consequences of multicollinearity for the estimates of the regression parameters and their estimated covariance matrix. The case where one eigenvalue (λ_m) is relatively very small is particularly simple. Consider the linear regression model

$$y = x\beta + \epsilon$$

where y is an n-vector containing the observations on the variable to be explained, X and $n \times m$ matrix, as before, containing n observations on each of m explanatory variables, β a vector of unknown parameters to be estimated and $\in$ a vector of disturbances, with zero means and covariance matrix $\sigma^2 I$. Let A denote the matrix of eigenvectors of $X' X$ and A the diagonal matrix containing the corresponding eigenvalues, then we have $X' X = A\ A\ A'$ with $A'\ A = I$. Then the inverse satisfies $(X' X)^{-1} = A\ A^{-1}\ A'$ and v_{ii}, the ith diagonal element of the covariance matrix $V = \sigma^2 (X' X)^{-1}$ can be written as

$$\upsilon_{ii} = \sigma^2 \sum_j a_{ij}^2 (1/\lambda_j)$$

where a_{ij} is the typical element of A. So v_{ii} is small if the a_{ij}^2 that correspond to small values of λ_j are small and large in the opposite case. This knowledge is helpful in understanding the problem of multicollinearity. Fomby et al. (1984) give a more extensive treatment and more references.

The next question is whether the relationship between principal components and multicollinearity can be exploited in order to solve the problems created by multicollinearity. It has been suggested that one might delete a number of principal components. Since the possibility exists that some of the principal components with small variances have a strong influence on the variable to be explained, it cannot be guaranteed that deleting these is a good choice. This may be decided by means of a preliminary test.

When applying the principal components method we transform a set of variables into linear combinations that are uncorrelated. Theil (1976) extends this approach in the context of the Rotterdam model of consumer demand. He constructs linear combinations of commodities that are *preference independent* and, hence, have a diagonal matrix of price coefficients in his demand system. In an example (p. 287) he transforms beef, pork and chicken into artificial preference independent commodities called inexpensive meat, beef/pork contrast and antichicken. He also gives an example containing clothing, footwear and other goods. His discussion on pages 311 and 312 is illustrative for the interpretation problems that may arise.

In general, one may say that principal components have elegant mathematical properties, but that their interpretation in applications is often far from simple.

See Also

► Factor Analysis

Bibliography

Amemiya, T. 1966. On the use of principal components of independent variables in two-stage leastsquares estimation. *International Economic Review* 7: 283–303.

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

Dhrymes, P.J. 1970. *Econometrics*. New York: Harper & Row.

Fisher, F.M. 1965. The choice of instrumental variables in the estimation of economy-wide econometric models. *International Economic Review* 6: 245–274.

Fomby, T.B., R.C. Hill, and S.R. Johnson. 1984. *Advanced econometric methods*. New York: Springer.

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

Kloek, T., and L.B.M. Mennes. 1960. Simultaneous equations estimation based on principal components of predetermined variables. *Econometrica* 28: 45–61.

Stone, J.R.N. 1947. On the interdependence of blocks of transactions. *Journal of the Royal Statistical Society, Series B* 9(Supplement): 1–45.

Theil, H. 1971. *Principles of econometrics*. Amsterdam/New York: North-Holland/Wiley.

Theil, H. 1976. *Theory and measurement of consumer demand*, vol. 2. Amsterdam: North-Holland.

Theil, H. 1983. Mathematical and statistical methods in econometrics. In *Handbook of econometrics*, vol. 1, ed. Z. Griliches and M.D. Intriligator. Amsterdam: North-Holland.

Thurstone, L.L. 1931. Multiple factor analysis. *Psychological Review* 38: 406–427.

Prisoner's Dilemma

Anatol Rapoport

The game nicknamed 'prisoner's dilemma' by A.W. Tucker has attracted wide attention, doubtless because it has raised doubts about the universal applicability of the so called Surething Principle as a principle of rational decision.

The game is illustrated by the following anecdote. Two men, caught with stolen goods, are suspected of burglary, but there is not enough evidence to convict them of that crime, unless one or both confess. They could, however, be convicted of possession of stolen goods, a lesser offence.

The prisoners are not permitted to communicate. The situation is explained to each separately. If both confess, both will be convicted of burglary and sentenced to two years in prison. If neither confesses, they will be convicted of possession of stolen goods and given a sixmonth prison sentence. If only one confesses, he will go scot-free, while the other, convicted on the strength of his partner's testimony, will get the maximum sentence of five years.

It is in the interest of each prisoner to confess. For if the other confesses, confession results in a two-year sentence, while holding out results in a five-year sentence. If the other does not confess, holding out results in a six-month sentence, while confession leads to freedom. Thus, 'to confess' is a *dominating strategy*, one that results in a preferred outcome regardless of the strategy used by the partner. A dominating strategy can be said to be dictated by the Sure-thing Principle. Nevertheless, if both, guided by the Sure-thing Principle, confess, both are worse off (with a two-year sentence) than if they had not confessed and had got a six-month sentence.

In this way, Prisoner's Dilemma is seen as an illustration of the divergence between individual and collective rationality. Decisions that are rational from the point of view of each individual may be defective from the point of view of both or, more generally, all individuals in decision situations where each participant's decision affects all participants.

Generalized to more than two participants (players), Prisoner's Dilemma becomes a version of the so called Tragedy of the Commons (Hardin 1968) It is in each farmer's interest to add a cow to his herd grazing on a communal pasture. But if each farmer follows his individual interest, the land may be overgrazed to everyone's disadvantage. Over-harvesting in pursuit of profit by each nation engaged in commercial fishing is essentially Tragedy of the Commons in modern garb.

Many social situations are characterized by a similar bifurcation between decisions prescribed by individual and collective rationality. Price wars and arms races are conspicuous examples. In the context of Prisoner's Dilemma, holding out would be regarded as an act of cooperation (with the partner, of course, not with the authorities); confession with noncooperation or defection.

Because the prescriptions of individual and collective rationality are contradictory, a normative theory of decision in situations of this sort becomes ambivalent. Attention naturally turns to the problem of developing a *descriptive* theory, one which would purport to describe (or to predict, if possible) how people, faced with dilemmas of this sort, actually decide under a variety of conditions.

As experimental social psychology was going through a rapid development in the 1950s, Prisoner's Dilemma became a favourite experimental tool. It enabled investigators to gather large masses of data with relatively little effort. Moreover, the data were all 'hard', since the dichotomy between a cooperative choice in a Prisoner's Dilemma game (C) and a defecting one (D) is unambiguous. Frequencies of these choices became the principal dependent variables in experiments on decision-making involving choices between acting in individual or collective interest. As for the independent variables, these ranged over the personal characteristics of the players (sex, occupation, nationality, personality profile), conditions under which the decisions were made (previous experience, opportunities for communication), characteristics or behaviour of partner, the payoffs associated with the outcomes of the game, etc. (cf. Rapoport et al. 1976, chs 9, 15, 18, 19).

Prisoner's Dilemma is usually presented to experimental subjects in the form of a 2 × 2 matrix, whose rows, C_1 and D_1, represent one player's choices, while the columns, C_2 and D_2 represent the choices of the other. The choices are usually made independently. Thus, the four cells of the matrix correspond to the four possible outcomes of the game: C_1C_2, C_1D_2, D_1C_2 and D_1D_2. Each cell displays two numbers, the first being the payoff to Row, the player choosing between C_1

and D_1, the second the payoff to Column, who chooses between C_2 and D_2. The magnitudes of the payoffs are such that strategy (choice) D of each player dominates strategy C. The decision problem is seen as a dilemma, because both players prefer outcome C_1C_2 to D_1D_2; yet to choose C entails forgoing taking advantage of the other player, should he choose C, or getting the worst of the four payoffs, should he choose D.

The experiments are usually conducted in one of three formats: (1) single play, where each player makes only one decision; (2) iterated play, in which several simultaneous sequential decisions are made by a pair of players; (3) iterated play against a programmed player, where the subject's co-player's choices are determined in a prescribed way, usually dependent on the subject's choices.

The purpose of a single play is to see how different subjects will choose when there is no opportunity of interacting with the other player. The purpose of iterated play with two bona fide subjects is to study the effects of interaction between the successive choices. The purpose of play against a programmed player is to see how different (controlled) strategies of iterated plays influence the behaviour of the subject, whether, for example, cooperation is reciprocated or exploited, whether punishing defections has 'deterrent' effect, etc. For an extensive review of experiments with a programmed player, see Oskamp 1971.

The findings generated by experiments with Prisoner's Dilemma are of various degrees of interest. Some are little more than confirmations of common sense expectations. For example, frequencies of cooperative choices in iterated plays vary as expected with the payoffs associated with the outcomes. The larger the rewards associated with reciprocated cooperation or the larger the punishments associated with double defection, the more frequent are the cooperative choices. The larger the punishment associated with unreciprocated cooperation, the more frequent are the defecting choices, and so on. As expected, opportunities to communicate with the partner enhance cooperation; inducing a competitive orientation in the subjects inhibits it.

Of greater interest are the dynamics of iterated play. Typically, the frequency of cooperative choices averaged over large numbers of subjects at first decreases, suggesting disappointment with unsuccessful attempts to establish cooperation. If the play continues long enough, average frequency of cooperation eventually increases, suggesting establishment of a tacit agreement between the players. The asymptotically approached frequency of cooperation represents only the mean and not the mode. Typically, the players) 'lock in' either on the C_1C_2 or on the D_1D_2 outcome (Rapoport and Chammah 1965).

Bimodality is observed also in iterated plays against a programmed player who cooperates unconditionally. Roughly one half of the subjects have been observed to reciprocate this cooperation fully, while one half have been observed to exploit it throughout, obtaining the largest payoff.

Comparison of the effects of various programmed strategies in iterated play showed that the so called Tit-for-tat strategy was the most effective in eliciting cooperation from the subjects. This strategy starts with C and thereafter duplicates the co-player's choice on the previous play. Of some psychological interest is the finding that the subjects are almost never aware that they are actually playing against their own mirror image one play removed. In a way, this finding is a demonstration of the difficulty of recognizing that others' behaviour towards one may be largely a reflection of one's behaviour towards them. Escalation of mutual hostility in various situations may well be a consequence of this deficiency.

Perhaps the most interesting result of Prisoner's Dilemma experiments with iterated play is that even if the number of iterations to be played is known to both subjects, nevertheless a tacit agreement to cooperate is often achieved. This finding is interesting because it illustrates dramatically the deficiency of prescriptions based on fully rigorous strategic reasoning.

At first thought, it seems that a tacit agreement to cooperate is rational in iterated play, because a defection can be expected to be followed by a retaliatory defection in 'self defence', so to say, by the other player with the view of avoiding the worst payoff associated with unreciprocated cooperation. However, this argument does not apply to the play known to be the last, because no retaliation can follow. Thus, D dominates C on the last play, and according to the Sure-Thing Principle, D_1D_2 is a foregone conclusion. This turns attention to the next-to-thelast play, which now is in effect, the 'last play', to which the same reasoning applies. And so on. Thus, rigorous strategic analysis shows that the strategy consisting of D's throughout the iterated play is the only 'rational one', regardless of the length of the series.

The backward induction cannot be made if the number of iterations is infinite or unknown or determined probabilistically. In those cases, provided the probability of termination is not too large, the 100 per cent D strategy is not necessarily dictated by individual rationality. The question naturally arises about the relative merit of various strategies in iterated play of Prisoner's Dilemma. This question was approached empirically by Axelrod (1984).

Persons interested in this problem were invited to submit programmes for playing iterated Prisoner's Dilemma 200 times. Each programme was to be matched with every other programme submitted, including itself. The programme with the largest cumulated payoff was to be declared the winner of the contest.

Fifteen programmes were submitted, Tit-for-tat among them. It obtained the highest score. A second contest was announced, this time with probabilistic termination, 200 iterations being the expected number. The results of the first contest together with complete descriptions of the programmes submitted were publicized with the invitation to the second contest. This time 63 programmes were submitted from six countries. Tit-for-tat was again among them (submitted by the same contestant and by no other) and again obtained the highest score.

The interesting feature of this result was the fact that Tit-for-tat did not 'beat' a single programme against which it was pitted. In fact, it cannot beat any programme, since the only way to get a higher score than the co-player is to play more D's than he, and this, by definition, Tit-for-tat cannot do. It can only either tie or lose, to be sure by no more than one play. It follows that Tit-for-tat obtained the highest score, because other programmes, presumably designed to beat their opponents, reduced each other's scores when pitted against each other, including themselves. The results of these contests can be interpreted as further evidence of the deficiency of strategies based on attempts to maximize one's individual gains in situations where both cooperative and competitive strategies are possible. Moreover, the superiority of cooperative strategies does not necessarily depend on opportunities for explicit agreements.

Support for the latter conjecture came from a somewhat unexpected source, namely, applications of game-theoretic concepts in the theory of evolution (Maynard Smith 1982; Rapoport 1985). Until recently, game-theoretic models used in theoretical biology were so called games against nature (e.g. Lewontin 1961). A 'choice of strategy' was represented by the appearance of a particular genotype in a population immersed in a stochastic environment. Degree of adaptation to the environment was reflected in relative reproductive success of the genotype, i.e. statistically expected numbers of progeny surviving the reproductive age. In this way, the population evolved towards the best adapted genotype.

In this model, adaptation depends only on the probability distribution of the states of nature occurring in the environment (e.g. wet or dry seasons) but not on the fraction of the population that has adopted a given strategy. When this dependence is introduced, the model becomes a genuine game-theoretic model with more than one bona fide player.

The model suggested by Prisoner's Dilemma appeared in theoretical biology in connection with combats between members of the same species, for example over mates or territories. Assuming for simplicity two modes of fighting, fierce and mild, we can see the connection to Prisoner's Dilemma by examining the likely result of evolution. In an encounter between a fierce and a mild fighter, the former wins, the latter loses. However, an encounter between two fierce fighters may impose more severe losses on both than an encounter between two mild fighters. With proper rank ordering or payoffs (relative reproductive success), the model becomes a Prisoner's Dilemma. Development of non-lethal weapons, such as backward curved horns or behavioural inhibitions may have been results of natural selection which made lethal combats between members of the same species rare.

Iterated combats suggest comparison of the effectiveness of strategies in iterated play. Maynard Smith and Price (1973) observed a computer-simulated population of iterated Prisoner's Dilemma players, using different strategies, whereby the payoffs were translated into differential reproduction rates of the players using the respective strategies. In this way, the) 'evolution' of the population could be observed. Eventually, the 'Retaliators', essentially Tit-for-tat players, replaced all others.

A central concept in game-theoretic models of evolution is that of the evolutionarily stable strategy (ESS). It is stable in the sense that a population consisting of genotypes representing that strategy cannot be 'invaded' by isolated mutants or immigrants, since such invaders will be disadvantaged with respect to their reproductive success. It has been shown by computer simulation that a population represented by programmes submitted to the above-mentioned contests evolved towards Tit-for-tat as an evolutionarily stable strategy. It was, however, shown subsequently that it is not the only such strategy.

In sum, the lively interest among behavioural scientists and lately many biologists in Prisoner's Dilemma can be attributed to the new ideas generated by the analysis of that game and by results of experiments with it. The different prescriptions of decisions based on individual and collective rationality in some conflict situations cast doubt on the very meaningfulness of the facile definition of 'rationality' as effective maximization of one's own expected gains, a definition implicit in all manners of strategic thinking, specifically in economic, political, and military milieus. Models derived from Prisoner's Dilemma point to a clear refutation of a basic assumption of classical economics, according to which pursuit of self-interest under free competition results in collectively optimal equilibria. These models also expose the fallacies inherent in assuming the) 'worst case' in conflict situations. The assumption is fully justified in the context of two-person zero sum games but not in more general forms of conflict, where interests of participants partly conflict and partly coincide. Most conflicts outside the purely military sphere are of this sort.

Finally, Prisoner's Dilemma and its generalization, the Tragedy of the Commons, provide a rigorous rationale for Kant's Categorical Imperative: act in the way you wish others to act. Acting on this principle reflects more than altruism. It reflects a form of rationality, which takes into account the circumstance that the effectiveness of a strategy may depend crucially on how many others adopt it and the fact that a strategy initially successful may become self-defeating *because* its success leads others to imitate it. Thus, defectors in Prisoner's Dilemma may be initially successful in a population of cooperators. But if this success leads to an increase of defectors and a decrease of cooperators, success turns to failure. Insights of this sort are of obvious relevance to many forms of human conflict.

See Also

▶ Game Theory
▶ Repeated Games

Bibliography

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

Hardin, G. 1968. The tragedy of the commons. *Science* 162: 1243–1248.

Lewontin, R.C. 1961. Evolution and the theory of games. *Journal of Theoretical Biology* 1: 382–403.

Maynard Smith, J. 1982. *Evolution and the theory of games*. Cambridge: Cambridge University Press.

Maynard Smith, J., and G.R. Price. 1973. The logic of animal conflict. *Nature* 246: 15–18.

Oskamp, S. 1971. Effects of programmed strategies on cooperation in the Prisoner's Dilemma and other mixed-motive games. *Journal of Conflict Resolution* 15: 225–259.

Rapoport, A. 1985. Applications of game-theoretic concepts in biology. *Bulletin of Mathematical Biology* 47: 161–192.

Rapoport, A., and A.M. Chammah. 1965. *Prisoner's Dilemma*. Ann Arbor: University of Michigan Press.

Rapoport, A., M. Guyer, and D. Gordon. 1976. *The 2 × 2 game*. Ann Arbor: University of Michigan Press.

Privacy, The Economics of

Grazia Cecere[1], Fabrice Le Guel[2], Matthieu Manant[2] and Nicolas Soulié[2]
[1]Telecom Ecole de Management, Institut Mines Telecom, Convergence Institut I2 Drive, Evry, France
[2]RITM Université Paris-Sud, Université Paris-Saclay, Convergence Institut I2 Drive, Sceaux, France

Abstract

The increasing digitalization of the economy and advances in data processing have drawn economists' attention to the role of personal data in markets. The economics of privacy aims to analyze how individuals, firms and policymakers interact in markets where personal data play a key role. The complexity of these markets is challenging for academics who rely on different disciplines and methods to investigate privacy issues, including field experiments. In this entry, we review this literature, and point out the need to take account of the complex interactions between the different economic agents, and highlight specific strategies regarding privacy issues. First, individuals might face a puzzling tradeoff between sharing data in order to access customized services and, on the other hand, protecting their personal data against potential data misuses. Second, industrial organizations and the literature of marketing study how firms can fit personal data into their strategies, and how this can spur new business models. Third, privacy regulation is being challenged as it aims both at protecting individuals' privacy while preserving firms' capacity to innovate. In this context, the main difficulty for policymakers is to shape a clear framework for both individuals and firms. Lastly, we bring attention to the need for further research to investigate the role of privacy as a business differentiator: in other words to establish a clear link between consumers' demand and firms' offer.

Keywords

Privacy; Regulation; Discrimination; Information asymmetry; Personalized advertising

JEL

D18; D82; D83; K2; L5; M31; M37

Introduction

The role of personal data in economics has been spurred on by the increasing digitalization of the economy, as it is now possible to collect, store and process huge amounts of data. In the economics literature, the concealing of individuals' personal information was first supposed to lead to information asymmetries between employee and employer, insured and insurer, etc., and thus, to possible market inefficiencies (see Stigler 1980; Posner 1981). However, especially in the context of internet economics, little is known about what might be the right balance for individuals between disclosure and protection of their personal information. For example, individuals might know (or not know) that firms may exploit their data for marketing purposes, or for discrimination in the job market. Recently, large parts of the economics and marketing literature have begun focusing on individual behaviours in different contexts (Acquisti et al. 2012), and in particular, it is investigating the effectiveness of firms' personalized advertising (Lambrecht and Tucker 2013) and the impact of privacy regulation (Campbell et al. 2015).

The literature provides numerous definitions and conceptions of privacy. For instance, Hirshleifer (1980) considers that privacy is the individual capability to act autonomously and independently of other individuals' control. In the same way, Westin (1967) defines privacy as the ability of individuals to control access to, and use of, their personal data, which would appear more relevant in the context of the internet. From an economics perspective, personal data have the characteristics of an information good that is non-rivalrous and non-excludable. Consequently, the secondary use of individual personal data by companies can occur without the individual's awareness. This can lead to possible negative externalities, particularly in the internet era. While the exploitation of personal data enables companies to propose better offers to users by reducing their search costs, it can also lead to discrimination or unwanted and inappropriate solicitations. The economics of privacy examines the policy implications related to the interactions between individuals' decisions to disclose personal data and firms' strategies to implement innovation using these data. The pervasiveness of internet economics (including the internet of things) in various sectors and industries has increased the importance of data to any kind of business activity, and in particular, in the health sector and when dealing with vulnerable populations such as children or people with disabilities.

The economics of privacy encompasses a range of disciplines such as industrial organization, marketing, behavioural economics, information systems, health economics, computer science and law. This strand of literature largely relies on the use of field experiments to study behaviours of both individuals and firms as it allows the assessment of causal inferences. Considering the contributions made by these disciplines, the present entry analyzes the emergence of new forms of complex interactions between the agents in the online market, namely consumers, firms and policymakers. The entry is organized as follows. Section "Individuals' Privacy Behaviours" investigates individuals' behaviours and individual awareness about privacy matters and identifies the role of information asymmetry between individuals and firms. It focuses on research in marketing and behavioural economics. Section "How Firms Use Personal Data" focuses on the market for privacy and firms' strategies (discrimination, hidden market, etc.). The data collected by companies can be used in different contexts, for example to inform individuals during advertising campaigns (e.g. of the best price at your favourite store) or conversely to discriminate against them during a job search. Section "Regulation and Policy Intervention" deals with the regulation of privacy, and it delves into analysis of the factors influencing the adoption of new modes of regulation, such as the spread of privacy-enhancing services. It also demonstrates how policymakers' analysis of privacy regulation is crucial since it provides the instruments to influence firms' strategies and to disseminate information (addressed) to individuals (Goldfarb and Tucker 2012b). It considers recent literature which suggests that major events such as Edward Snowden's revelations about possible misuses of personal data by firms drew the attention of individuals to privacy issues.

Individuals' Privacy Behaviours

The advent of the internet and the increasing adoption of connected devices mean that individuals currently face many more privacy issues than before. Understanding individuals' privacy concerns is crucial since they can hamper the diffusion of online services and, more generally, the growth of ICTs. Above all, there is an obvious tradeoff for users between sharing and protecting their personal data. On the one hand, sharing can be beneficial as it allows users access to personalized services. Individuals can also benefit from the externalities from other individuals' data disclosures. For instance, the (good or bad) recommendations about a product on a website, which reveal people's preferences, can benefit other users who have access to this information. On the other hand, users need to consider possible misuses of their data (Acquisti et al. 2016). Disclosing personal information is not the only way for individuals to have their data collected, since firms can use activity, behavioural or locational data to find out individuals' characteristics. As an example, Facebook Likes can help to predict a range of highly sensitive personal data, like political views or sexual orientation (Kosinski et al. 2013). Consequently, the rationale adopted by users to privacy issues is interesting from an economic point of view. From the perspective of a market for privacy, users may be aware that their personal traits and attributes (age, address, gender, revealed preferences via social media or purchases, comments, etc.) have an economic value for firms.

By default, many existing theoretical models consider users as having complete information about firms' strategies; in other words, they suppose that individuals know how firms might use their personal data, and how to react to these strategies. In fact, these models are related closely to the acceptance of the internet model by users. They deal with how markets would be affected where individuals are reluctant to share their data, and with the effectiveness of privacy protection strategies. While the earliest theoretical approaches consider hiding personal information to be a suspicious strategy, which may be inefficient from a social perspective (Hirshleifer 1980; Stigler 1980; Posner 1981), later models are more realistic and consider that this might be a rational strategy for consumers. Thus, Varian (1997) admits that individuals might find it beneficial not to reveal private information and introduces the idea that individuals may consider secondary usage in their privacy strategies. Within a Coasian perspective, Kahn et al. (2000) take into account explicitly the awareness of consumers able to "deal" in their private information, and able to evaluate the costs and benefits of their privacy choices. Subsequent models provide evidence of consumer strategies which firms need to consider in markets. Indeed, Fudenberg and Tirole (1998) show that a monopoly selling durable goods may employ different strategies according to the type of good, that is the type of consumer – anonymous, semi-anonymous or identified. In a duopoly setting, Villas-Boas (1999) provides evidence that consumers have an interest in revealing their preferences to competitors and to be patient, that is they care about the future, in order to get lower prices from competitors who try to attract them. Thus, theoretical models show that consumers' preferences drive the strategies of firms whose business models are based on the disclosure of personal data by users. In some way, these approaches question the consequences of the existence of asymmetric information between users and firms about firms' privacy strategies. Indeed, the literature points out that individuals may not know what firms or third parties intend to do with their data, and for that reason, may be reluctant to disclose them.

In order to capture the issue of information asymmetry, which is highlighted in theoretical models, and, more broadly, the way individuals reason about privacy, researchers propose empirical approaches and field experiments. The literature tries to understand the diversity of privacy behaviours and to provide evidence of biases which might influence privacy decision making.

Empirical research is interested in estimating users' privacy concerns by emphasizing the motivations and conditions which lead individuals to disclose (or not) personal information. Through an interdisciplinary review, Smith et al. (2011) identify five main factors that might explain an individual's level of concern about online privacy: privacy experience, privacy awareness, personality, demographics and culture. In the case of online services, consumers tend to disclose personal information to access personalized services (Chellappa and Sin 2005). Nevertheless, the literature provides evidence of a privacy paradox, or an inconsistency between what is said and what is done. These differences in privacy choices can be context dependent (Nissenbaum 2004) or can be due to intrinsic characteristics, that is individual preferences or cognitive biases. Field experiments show that personal data disclosure can be explained by access to immediate gratification, by incomplete information and by individuals' bounded rationality (Acquisti 2004; Acquisti and Grossklags 2005). Through a series of diverse field experiments, Acquisti et al. (2012) show that survey respondents disclose more personal data if other respondents do so – the herd effect – and that the level of disclosure increases if the questions progress from more to less intrusive. In a study of people's interactions, Forman et al. (2008) find that consumer-generated product reviews containing identity-descriptive information are rated more positively by community members and are associated with an increase in product sales. Moreover, information disclosure tends to become the norm and to lead other reviewers of a product to do the same.

The literature has underlined that individuals' privacy concerns increase once consumers have, or have heard about, bad privacy experiences, such as identity theft or unwanted secondary uses (Smith et al. 1996). The extent of their concern over privacy depends also on an individual's awareness of firms' privacy practices (Malhotra et al. 2004). In a field experiment, Marreiros et al. (2016) test whether privacy actions and attitudes can be influenced by exposure to information about the advantages and disadvantages associated with disclosing personal information online. The results suggest that privacy concerns emerge once users are asked to think about privacy issues.

From a market perspective, individuals' awareness of the potential risks associated with privacy abuses can have an impact on the demand for privacy. In a field experiment designed on a shopping search engine interface, Tsai et al. (2011) show that individuals' purchasing intentions increase if online retailers display salient information about privacy protection clearly. Acquisti et al. (2013) disentangle the issue of privacy valuations by suggesting that individuals value privacy more when they have it than when they do not. In other words, they highlight the existence of an endowment effect. Using a large dataset, Goldfarb and Tucker (2012b) use a refusal to disclose personal information (here individual income) to measure demand for privacy. They show that, overall, there is an increasing percentage of individuals who do not disclose personal data and that there is a generational pattern: younger individuals disclose more compared to older people.

Open Research Questions

Regarding individuals' privacy behaviours, there are several open research questions. For example, knowing the differing privacy concerns among users is essential in order to conceive adequate privacy regulation. A regulation aimed at users who are not concerned about privacy is likely to be ineffective. For this reason, it is of interest to investigate the younger generation's privacy concerns, and especially the privacy concerns of teenagers. Demand for privacy appears also to be an important feature, and there is a need to design field experiments to estimate the demand for privacy, that is if users find a good or a service with privacy characteristics more attractive.

How Firms Use Personal Data

The marketing and industrial organization literatures mostly investigate how companies exploit personal data, and how personal data can spur new business models and, thus, generate innovation. Big data, data analysis and progress in business analytics permit data to be retrieved and analyzed at an unprecedented level (Acquisti 2014). In a seminal contribution, Varian (1997) distinguishes between first and "second usage" of personal information by internet companies: first usage facilitates firms' interactions with customers, and second usage occurs when the firms pass on information to one or more other firms – "third parties" – better able to exploit personal data. This distinction defines a primary market involving customers and internet companies, and a secondary market involving internet companies and third parties.

In the primary market involving internet companies and customers, personal data is first used to design more effective advertising campaigns aimed at individuals and to set prices that are close to individuals' willingness to pay. Second, both the exploitation of clickstream data that provide detailed information on how individuals interact with websites and advertising and the increased importance of algorithms in internet economics help speed up the pace of innovation.

According to price discrimination theory, the link between firms' strategies and personal information, and thus privacy, is central (Taylor 2004). Exploiting personal information could facilitate first degree price discrimination or, more realistically, third degree discrimination as these data permit a company to identify an individual's reservation price. Consistent with a positive effect of discrimination for users, a recent theoretical work by Belleflamme and Vergote (2016) suggests that the use of technologies to conceal personal information might reduce consumer surplus.

For most online firms, personal ad revenue is a major source of income (Martin and Murphy 2016). However, empirical works based on large field experiments provide some puzzling results regarding the effectiveness of personalized ads. In particular, in a field experiment on a popular social media website, Tucker (2014) shows that the effectiveness of an ad increases if individuals have more control over their personal data. In another field experiment, Lambrecht and Tucker (2013) show that dynamic ad retargeting is, on average, less efficient than generic ads. Dynamic retargeting effectiveness increases once individuals have more information related to the products they want to buy. In addition to price discrimination, other forms of discrimination can arise, in particular, in the labour market where recruiters can discriminate among candidates on the basis of information available on social media (Acquisti and Fong 2015; Manant et al. 2016). The article by Lambrecht and Tucker (2016) adds to this literature the role of social media algorithms which reproduce offline discrimination of individuals and particularly, that of women. Overall, these articles highlight how firms can exploit personal data which users leave online.

Data exploitation strategies suppose that firms can access individuals' personal data. This is why the theoretical literature highlights the need for firms to take account of individuals' privacy concerns. Taylor (2004) shows that firms can employ different strategies depending on the privacy regulation and consumers' expectations: firms prefer a disclosure regime if consumers are naïve and a confidential regime if consumers expect that their personal data will be used by firms. Acquisti and Varian (2005) provide similar results – that is that firms' strategies rely on consumers' preferences for personalized services. This need to take account of consumers' strategies to protect their privacy is confirmed by dynamic approaches (Villas-Boas 2004; Armstrong and Zhou 2010). Internet companies largely rely on the distribution of services or goods in exchange for users' registration. The effectiveness of these practices is not straightforward. In a theoretical model, Morath and Münster (2017) show that a monopoly firm can benefit from ex-ante registration requirements, in particular, if future purchases are considered. To influence consumers' decisions to register, consumer discounts seem appropriate and allow consumer surplus to be increased.

In the market involving internet companies and third parties, firms can also exploit personal data by selling it. While there are many marketing companies such as BlueKai and Avarto which specialize in data management, there is a need to understand the functioning of personal data markets and the role of the companies in these markets. Secondary use of personal information arises if data are sent to third parties or data brokers, that is to data aggregators, advertisers, or, more broadly, to competent departments within a firm (Akçura and Srinivasan 2005). Third party use and secondary use of personal data within the same firm seem to be less legitimate if personal data are sent without the awareness of the user. Taylor (2004) considers that this behaviour is welfare-diminishing for consumers. Using a theoretical model, Akçura and Srinivasan (2005) also show that this secondary market can result in a dramatic decrease in consumer welfare. The existence of the market for personal data implicitly questions the value of data to firms (Spiekermann et al. 2015). From this perspective, industrial organization scholars first may have to assess whether personal data are a good per se. Farrell (2012) contributes to this discussion by showing that personal data can be considered as a good and that privacy protection can be seen as a strategic parameter for profit maximizing firms. In this context, the contribution by Kummer and Schulte (2016) is relevant since it delves into the business models of smartphone applications by highlighting a tradeoff between price and personal data for both the market's supply and the demand side, seeing personal data as an alternative business model for free services. An OECD (2013) report provides details of different methods to evaluate personal data and offers some insights into the firms operating in the market involving internet companies and third parties.

Open Research Questions

There remain various research questions related to the industry structure of the personal data market and the extent to which personal data are part of the business models of internet companies. This is particularly relevant in the market for smartphone applications, where an increasing number of free applications are available. An important issue is to see if privacy can be a business differentiator.

Regulation and Policy Intervention

The emergence of new businesses based on personal data drew the regulators' attention to the need to find the right balance between protecting privacy and promoting data sharing to encourage innovation and improve services. Formulating public intervention in privacy is a complex issue. First, the innovation in sectors where personal data play a key role is challenging and competition among players is high. Second, the design of privacy regulation can affect the behaviours of both individuals and firms. Indeed, market interactions, namely those between firms and consumers, but also between firms and third party companies, can lead to unexpected consequences, which can trigger the effectiveness and the evaluation of the policy. Moreover, while privacy regulation is directed towards consumers and firms, it can also have indirect consequences on market structure. All in all, the direction and the size of these effects are unclear. While regulation helps to create the premise of a clear framework for companies, there is a need to understand the overall role of personal data on the markets. For this reason, we focus here on the principles that govern privacy regulation and on theoretical and empirical evidence of the impact of this regulation on the markets, but also on how privacy-enhancing technologies and data breaches can shape these markets.

Regulation and Self-Regulation

Focusing on the instruments used by the regulatory authorities helps understand how privacy regulation can intervene in the markets. In the USA where the Federal Trade Commission (FTC) provides guidelines at the sectoral level, self-regulation prevails. The main goal of self-regulation is to stimulate "competition on privacy," and then alleviate market failures. This approach considers both sides of the market by assuming that consumers can make decisions to enact their privacy preferences, and that companies are supposed to respect a principle of transparency and control over privacy issues, for example, by giving detailed information on data collection as well as on the use of data. In this perspective, privacy policies rely on a "notice-and-consent" principle where individuals are supposed to read privacy policies and consent or not to the terms of service (Cranor 2012). Privacy policies are expected to provide information to individuals on firms' practices about how their personal data are gathered, used, shared and secured (Marotta-Wurgler 2016). However, empirical evidence shows that those policies are too long to read and too complex to understand for a non-practitioner (McDonald and Cranor 2008). In line with the self-regulation approach, FTC policy has also encouraged the creation of third party certification services and online seals such as TRUSTes and BBBs to help decrease individuals' cognitive costs of assessing the eventuality of potential privacy threats. Nevertheless, adverse selection can emerge with such private seals. In particular, empirical research has shown that websites certified by TRUSTe are more than twice as likely to be untrustworthy as uncertified sites (Edelman 2011). In terms of policy implications, it suggests that regulatory intervention is necessary to ensure the quality of private seals.

In Europe, the regulatory approach is focused more on providing a general framework to protect consumers across sectors, which is substantially different from the self-regulation approach. Regulation aims at alleviating the adverse selection problem by ensuring more guarantees for individuals and a stringent environment for companies. The current debate was triggered in 2016 by the publication of the Data Protection Directive updating the previous Data Protection Directive 95/46/EC Directive and ePrivacy 02/58/EC Directive. This new framework allows individuals access to more information about how their data are processed by companies and gives individuals the right to have their data forgotten – that is, they can ask for deletion or modification of their data by the data holders. The Directive promotes the use of "privacy by design" which aims to embed privacy in the very early phases of the development of a product or a service, as well as throughout its development. Overall, while the European approach is supposed to bring more transparency and stronger protection for individuals than in the USA, it must also be more costly for firms that comply with, since it imposes stronger obligations on them.

The evolving practices of regulation dealing with privacy issues have also had an impact on the commercial relations between the USA and Europe. As an illustration, Snowden's revelations of mass surveillance programs have encouraged the replacement of the previous US-EU Safe Harbor Agreement by the EU-US Privacy Shield, which took effect in July 2016. This agreement pushes for more cooperation between US and European Data Protection Authorities. This obliges US companies to ensure transparency about trans-frontier transfers of personal data and stronger protection of personal data.

Impacts of Privacy Regulation

Since privacy regulation affects both individuals and firms, the literature has identified different levels of impact of this regulation, that is, the impact on agents' behaviours, but also on market structure.

First, the literature shows that the different principles of privacy regulation have direct effects on individuals' choices. Looking at the variation in US State genetics privacy laws over time, Miller and Tucker (2015) show that giving individuals control over the redisclosure of their genetic tests by hospitals encouraged the diffusion of this practice, while requiring informed consent deterred individuals from undertaking genetic testing services. In a large-scale field experiment, Goldfarb and Tucker (2011) study how the enactment of the European privacy regulation has impacted the effectiveness of banner ads on individual behaviours. They show that the intention to buy has decreased since EU regulation restricted the use of data related to customers' past browsing behaviours.

The increased use of open data and mass data collection can also influence consumer behaviours and have potential commercial outcomes. Marthews and Tucker (2015) show that Edward Snowden's revelations about the US government surveillance changed the type of requests conducted on Google Search. The result suggests that this event has affected individual behaviours by increasing the demand for privacy.

Second, while privacy regulation is directed towards consumers and firms, it can also have indirect consequences on the market structure. Campbell et al. (2015) propose a theoretical model to show that a consent-based approach, even if it deters consumers and imposes costs on all firms, may disproportionately benefit generalist firms that offer a large scope of services, rather than specialist firms. This regulation regime thus affects the competitive structure of the market. In the context of the impact of sectoral privacy regulation on organizations' activities, there is a large literature studying the impact of health regulation in the USA on the diffusion of hospital information technologies. Exploiting the variation in State privacy legislation in the USA, Miller and Tucker (2009) show that privacy protection can hamper the adoption of these technologies by hospitals if they are unable to take advantage of patient information from other hospitals. US privacy law restricts cases where hospitals can exchange patient data. The regulation can also have an impact on the location of internet companies. Using a sample of the most visited websites worldwide, Rochelandet and Tai (2016) demonstrate empirically that internet firms prefer to be located in countries where data collection and exploitation are less regulated. Inversely, tax instruments can help reduce data collection by internet platforms, while an opting-out option, where users can access the platform with no data collection, induces the platform to raise the level of data exploitation (Bloch and Demange 2016).

The increased importance of data has also stimulated mergers and acquisitions aiming to increase their competitive advantages. These practices might radically change the structure of the market. On the one hand, they can counterbalance the internet giants currently in place, and on the other hand, they can create the conditions for anti-competitive data-driven strategies once these operations aim to prevent rivals from accessing data or hamper the access of consumers to competitors' platforms (Goldfarb and Tucker 2012a). As an example, the Google/DoubleClick merger has illustrated how the standard tools of competition policy cannot really assess consumer privacy issues since privacy issues are related to non-price attributes. In this respect, the role of data brokers can be central. In a theoretical model, Clavorà Braulin and Valletti (2016) show that it is possible to achieve first best allocation only when data are sold non-exclusively. When a data broker sells the data exclusively, this creates inefficient allocations.

Privacy-Enhancing Services: Privacy as a "Business Differentiator"

The increasing level of individuals' privacy awareness can have an impact on demand for privacy, which then generates conditions conducive to the adoption of privacy-enhancing services. After the Snowden revelations, the number of DuckDuckGo service users, a search engine that does not register users' IP addresses, sharply increased by about 600% (Wired 2017). In January 2017, it had more than 14 million searches. In this respect, the protection of personal data can represent a differentiation strategy for internet companies, "pushing to the top" the most privacy respectful firms. In such a case, the decrease in personal data collection would offset the increase in the number of consumers becoming more confident. An important contribution in this respect is the theoretical model of Casadesus-Masanell and Hervas-Drane (2015) showing that higher competition intensity in the marketplace need not improve privacy when consumers exhibit low willingness to pay. However, so far there is no clear evidence that people do understand the current model of free internet services in which they give personal data in exchange for free services, and that they might be better off paying for it and protecting their privacy. Privacy-enhancing services can remain a niche market if demand for these services does not increase, which Farrell (2012) defines as a "dysfunctional equilibrium". In this respect, moving from free services to paid services is a major challenge in the future for privacy-enhancing services. Nowadays, the large majority of these services are freely available (examples are blogs, website contents or smartphone applications). The main sources of income of these internet services are advertising, e-shopping and personal data (Lambrecht et al. 2014). The advance of services that help block advertising, as well as the increase of individuals' privacy concerns, might challenge the model of free web services (e.g. Adblock Plus, Google Contributor). In particular, the demand for privacy can influence the willingness to pay for services which are respectful of privacy.

Open Research Question

With the "internet of things" generating huge amounts of data ("big data"), and then the proliferation of algorithms that implement artificial intelligence and machine learning, many choices can be suggested to individuals. This could have a positive or negative impact on social welfare. Potential discrimination due to these technologies will require special attention from the regulator but also innovations on their part.

Further Developments

We identify the existence of a privacy market linking users' demand for privacy. Looking at the drawbacks of privacy policy, privacy-enhancing technologies (PET) can be seen as an alternative. An example was the P3P project (Platform for Privacy Preferences Project) which was dedicated to the creation of machine-readable privacy policies, aimed to normalize privacy policies and help users to better understand them and increase individuals' trust. However, these solutions have also failed, due partly to their complexity. More advanced solutions, such as encryption, differential privacy or decentralized personal data systems, based on the blockchain technology, have been suggested to help build a market for personal data, but these technologies do not yet guarantee a possible re-identification of the personal data owner (Zyskind et al. 2015).

Starting from these weaknesses, and by simplifying the signal to its maximum, behavioural economists and psychologists suggest that information related to privacy issues should be conveyed in a simple and standardized manner in order to ensure that consumers understand it (Bhargava and Loewenstein 2015). Nudging privacy seems to be a promising practice that is complementary to effective regulation and individuals' empowerment (Lazer et al. 2009).

As a summary, it appears to be particularly complicated to achieve an appropriate balance between information sharing and information hiding. Therefore, there is no single way to reduce information asymmetry. Regulation, market-based or technologies-based solutions can be seen as complementary. In this perspective, privacy protection as a business differentiator "pushing to the top" the most privacy respectful firms deserves consideration, although it depends *in fine* on consumers' preferences for privacy.

See Also

- ▶ Advertising
- ▶ Cryptocurrency
- ▶ Field Experiments
- ▶ Internet and the Offline World
- ▶ Online Platforms, Economics Of
- ▶ Pricing Services Online, Economics Of

Bibliography

Acquisti, A. 2004. Privacy in electronic commerce and the economics of immediate gratification. In *Proceedings of the 5th ACM conference on electronic commerce*, 21–29. New York: ACM.

Acquisti, A. 2014. From the economics of privacy to the economics of big data. In *Privacy, big data, and the public good: Frameworks for engagement*, ed. S. Bender, J. Lane, H. Nissenbaum, and V. Stodden, 76–95. New York: Cambridge University Press.

Acquisti, A., and C. Fong. 2015. *An experiment in hiring discrimination via online social networks*. dx.doi.org/10.2139/ssrn.2031979.

Acquisti, A., and J. Grossklags. 2005. Privacy and rationality in decision making. *IEEE Security and Privacy* 3 (1): 26–33.

Acquisti, A., and H.R. Varian. 2005. Conditioning prices on purchase history. *Marketing Science* 24 (3): 367–381.

Acquisti, A., L.K. John, and G. Loewenstein. 2012. The impact of relative standards on the propensity to disclose. *Journal of Marketing Research* 49: 160–174.

Acquisti, A., L.K. John, and G. Loewenstein. 2013. What is privacy worth? *The Journal of Legal Studies* 42 (2): 249–274.

Acquisti, A., C. Taylor, and L. Wagman. 2016. The economics of privacy. *Journal of Economic Literature* 54 (2): 442–492.

Akçura, M.T., and K. Srinivasan. 2005. Research note: Customer intimacy and cross-selling strategy. *Management Science* 51 (6): 1007–1012.

Armstrong, M., and J. Zhou. 2010. *Conditioning prices on search behaviour*, Munich Personal RePEc Archive Paper 19985.

Belleflamme, P., and W. Vergote. 2016. Monopoly price discrimination and privacy: The hidden cost of hiding. *Economic Letters* 149: 141–144.

Bhargava, S., and G. Loewenstein. 2015. Behavioral economics and public policy 102: Beyond nudging. *American Economic Review* 105 (5): 396–401.

Bloch, F., and G. Demange. 2016. *Taxation and privacy protection on internet platforms,* halshs-01381044.

Campbell, J., A. Goldfarb, and C. Tucker. 2015. Privacy regulation and market structure. *Journal of Economics and Management Strategy* 24 (1): 47–73.

Casadesus-Masanell, R., and A. Hervas-Drane. 2015. Competing with privacy. *Management Science* 61 (1): 229–246.

Chellappa, R.K., and R.G. Sin. 2005. Personalization versus privacy: An empirical examination of the online consumer's Dilemma. *Information Technology and Management* 6 (2): 181–202.

Clavorà Braulin, F., and T. Valletti. 2016. Selling customer information to competing firms. *Economics Letters* 149: 10–14.

Cranor, L.F. 2012. Necessary but not sufficient: Standardized mechanisms for privacy notice and choice. *Journal on Telecommunications and High Technology Law* 10: 273–307.

Edelman, B. 2011. Adverse selection in online "trust" certifications and search results. *Electronic Commerce Research and Applications* 10 (1): 17–25.

Farrell, J. 2012. Can privacy be just another good? *Journal on Telecommunications and High Technology Law* 10 (2): 251–261.

Forman, C., A. Ghose, and B. Wiesenfeld. 2008. Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets. *Information Systems Research* 19 (3): 291–313.

Fudenberg, D., and J. Tirole. 1998. Upgrades, tradeins, and buybacks. *RAND Journal of Economics* 29 (2): 235–258.

Goldfarb, A., and C.E. Tucker. 2011. Privacy regulation and online advertising. *Management Science* 57 (1): 57–71.

Goldfarb, A., and C.E. Tucker. 2012a. Privacy and innovation. *Innovation Policy and the Economy* 12 (1): 65–90.

Goldfarb, A., and C.E. Tucker. 2012b. Shifts in privacy concerns. *American Economic Review* 102 (3): 349–353.

Hirshleifer, J. 1980. Privacy: Its origin, function, and future. *The Journal of Legal Studies* 9 (4): 649–664.

Kahn, C.M., J. McAndrews, and W. Roberds. 2000. *A theory of transactions privacy.* Federal Reserve Bank of Atlanta Working Paper 2000–22.

Kosinski, M., D. Stillwell, and T. Graepel. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences* 110 (15): 5802–5805.

Kummer, M.E., and P. Schulte. 2016. *When private information settles the bill: Money and privacy in Google's market for smartphone applications*, ZEW-Centre for European Economic Research Discussion Paper, 16-031.

Lambrecht, A., and C.E. Tucker. 2013. When does retargeting work? Information specificity in online advertising. *Journal of Marketing Research* 50 (5): 561–576.

Lambrecht, A., and C.E. Tucker. 2016. *Algorithmic bias? An empirical study into apparent gender-based discrimination in the display of STEM career ads*. Social Science Research Network, WP.

Lambrecht, A., A. Goldfarb, A. Bonatti, A. Ghose, D. Goldstein, R. Lewis, A. Rao, N. Sahni, and S. Yao. 2014. How do firms make money selling digital goods online? *Marketing Letters* 25: 331–341.

Lazer, D., A. Pentland, L. Adamic, S. Aral, A.-L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, and M. Van Alstyne. 2009. Social science: Computational social science. *Science* 323 (5915): 721–723.

Malhotra, N.K., S.S. Kim, and J. Agarwal. 2004. Internet users' information privacy concerns (IUIPC): The construct, the scale, and a causal model. *Information Systems Research* 15 (4): 336–355.

Manant, M., S. Pajak, and N. Soulié. 2016. *Can social media lead to labour market discrimination: A field experiment*. WP.

Marotta-Wurgler, F. 2016. Self-regulation and competition in privacy policies. *The Journal of Legal Studies* 45 (S2): 13–39.

Marreiros, H., M. Tonin, and M. Vlassopoulos. 2016. *"Now that you mention it": A survey experiment on information, salience and online privacy* 140:1–17.

Marthews, A., and C.E. Tucker. 2015. *Government surveillance and internet search behavior*. Available at SSRN 2412564.

Martin, K.D., and P.E. Murphy. 2016. The role of data privacy in marketing. *Journal of the Academy of Marketing Science* 45:135–155.

McDonald, A.M., and L.F. Cranor. 2008. The cost of reading privacy policies. *I/S: A Journal of Law and Policy for the Information Society* 4 (3): 540–565.

Miller, A.R., and C. Tucker. 2009. Privacy protection and technology diffusion: The case of electronic medical records. *Management Science* 55 (7): 1077–1093.

Miller, A.R., and C. Tucker. 2015. *Privacy protection, personalized medicine and genetic testing*. dx.doi.org/10.2139/ssrn.2411230.

Morath, F. and J. Münster. 2017. Online shopping and platform design with ex ante registration requirements. *Management Science*. doi:org/10.1287/mnsc.2016.2595.

Nissenbaum, H. 2004. Privacy as contextual integrity. *Washington Law Review* 79 (1): 119–158.

Posner, R.A. 1981. The economics of privacy. *The American Economic Review* 71 (2): 405–409.

Rochelandet, R., and S. Tai. 2016. Do privacy laws affect the location decisions of internet firms? Evidence for privacy havens. *European Journal of Law and Economics* 42 (2): 339–368.

Smith, H.J., J.S. Milberg, and J.S. Burke. 1996. Information privacy: Measuring individuals' concerns about organizational practices. *MIS Quarterly* 20 (2): 167–196.

Smith, H.J., T. Dinev, and H. Xu. 2011. Information privacy research: An interdisciplinary review. *MIS Quarterly* 35 (4): 989–1015.

Spiekermann, S., A. Acquisti, R. Böhme, and K.L. Hui. 2015. The challenges of personal data markets and privacy. *Electronic Markets* 25 (2): 161–167.

Stigler, G.J. 1980. An introduction to privacy in economics and politics. *The Journal of Legal Studies* 9 (4): 623–644.

Taylor, C.A. 2004. Consumer privacy and the market for customer information. *RAND Journal of Economics* 35 (4): 631–665.

Tsai, J.Y., S. Egelman, L. Cranor, and A. Acquisti. 2011. The effect of online privacy information on purchasing behavior: An experimental study. *Information Systems Research* 22 (2): 254–268.

Tucker, C.E. 2014. Social networks, personalized advertising, and privacy controls. *Journal of Marketing Research* 51 (5): 546–562.

Varian, H.R. 1997. Economic aspects of personal privacy. In *Privacy and self-regulation in the information age*. Washington, DC: US Department of Commerce.

Villas-Boas, J.M. 1999. Dynamic competition with customer recognition. *RAND Journal of Economics* 30 (4): 604–631.

Villas-Boas, J.M. 2004. Price cycles in markets with customer recognition. *RAND Journal of Economics* 35 (3): 486–501.

Westin, A. 1967. *Privacy and freedom*. New York: Atheneum Publishers.

Wired. 2017. *DuckDuckGo hits 10 billion anonymous searches after big 2016*. http://www.wired.co.uk/article/duckduckgo-10-billion-searches. Last retrieved 30 Jan 2017.

Zyskind, G., O. Nathan, and A. Pentland. 2015. Enigma: Decentralized computation platform with guaranteed privacy. *arXiv preprint arXiv:1506.03471*.

Privatization

John Vickers

Abstract

Privatization is the transfer from government to private parties of the ownership of firms. Privatization programmes have been carried out worldwide since the mid-1980s, with important consequences for economic efficiency, public finance, and distribution. In competitive industries privatization generally has positive effects on incentives and performance. The economic consequences of privatizing firms with market power depend on the effectiveness of regulation and competition policy. These points are illustrated by experience in Britain, a leading exponent of privatization policies.

Keywords

Allocative efficiency; Bankruptcy; Barriers to entry; Bonds; Commitment; Competition law and policy; Competitive tendering; Control rights; Cross-subsidy; Duopoly policy; Expropriation; Firm, theory of; Hard budget constraint; Interconnection; Liberalization; Market power; Merger policy; Ownership; Present value; Price cap regulation; Price control; Principal and agent; Privatization; Public finance; Public–private partnerships; Rate-of-return regulation; Redistribution of wealth; Regulation of monopoly; Share options; State-owned enterprises; Takeovers; Universal service obligations

JEL Classifications

H0

'Privatization' is defined here as the transfer from government to private parties of the ownership of firms. This definition is not so broad as to embrace, for example, the sale of publicly owned housing and natural resources, contracting out the supply of publicly financed services, or the introduction of user charges for services previously provided at public expense. However, some of the economic principles for privatizing firms apply more generally.

This article is in two parts. The first part addresses some economic and financial principles of privatization, beginning with the basic question: how does ownership matter for economic efficiency? It is concluded that, at least for firms with significant market power, this question must be addressed in conjunction with the framework of regulation and competition that accompanies public or private ownership. The second part examines some aspects of privatization in practice, particularly in Britain, the leading exponent of the policy in the 1980s.

Privatization: Principles

Ownership and Economic Efficiency

If privatization is defined as the transfer of ownership, the first question is: what is ownership? According to the incomplete contracts view of the firm (see Hart 1995), ownership of an asset is to be identified with residual rights of control – rights to make decisions in the domain not already subject to contractual obligations. No such rights would exist in a world of complete contracts, where ownership, and hence privatization, would therefore be irrelevant.

The ultimate owners of sizable firms typically delegate the exercise of residual control rights to professional managers (whose identity may or may not be affected by privatization). Privatization affects principal–agent relationships between owners and managers by changing (*a*) the principals and hence their objectives, (*b*) the means of monitoring and giving incentives to the agents, and (*c*) the scope and incentives for action by the former public principals.

As to (*a*), a limitation to the economic theory of privatization is that there is no definitive theory of the firm under public ownership. In some sense the ultimate owners are the general public, but, even if their preferences could satisfactorily be aggregated into a welfare measure, it would be pious to suppose that government ministers or bureaucrats would necessarily exercise their authority over public firms to maximize welfare, avoiding distraction by political considerations, influence by well-organized vested interests, and so on. With private firms the usual assumption that owners seek to maximize profit or share value seems a tolerable approximation for present purposes, except perhaps if workers or consumers have large ownership stakes.

Since private, unlike public, ownership claims are generally tradable, privatization can alter the monitoring and incentives of managers by changing information conditions. For example, managers' rewards can be related to share price performance. In so far as share prices reflect the value of the firm, managers can thereby be given incentives to enhance firm value. Stock market investment analysts become a new source of managerial monitoring. However, free-rider considerations imply that monitoring by private owners might be limited, especially if share ownership is diffuse.

The tradability of ownership claims also means that privatized firms, unless they are given special protection, are potentially open to takeover threats, whereas publicly owned firms obviously are not. It is a matter for debate whether such threats from the market for corporate control are effective in disciplining managers of private firms to act in shareholder interests. Private firms also face the possibility of bankruptcy, in which case residual control rights shift to debt-holders.

Privatization changes the relationship between government and the firm. Thus public officials may lose power to intervene in the running of the firm. Moreover, and perhaps most important, the credibility of government commitment not to intervene may be enhanced by privatization so that, for example, managers face harder budget constraints and hence stronger incentives (see Schmidt 1996).

Nevertheless, privatization might not make government commitment not to intervene completely credible, especially if the firm remains subject to regulation or dependent on public subsidy. In any event, the government retains powers of taxation, and ultimately there is the possibility that privatization might be reversed, possibly on terms disadvantageous to private owners. To the extent that these factors give rise to risk of more or less subtle expropriation by government, private investment incentives may be adversely affected.

From the considerations above it follows that the consequences of privatization are likely to be influenced by the extent of market power enjoyed by the firm in question. For a firm that operates in competitive conditions, the shift from 'public' to profit objectives raises no concerns about the exercise of market power, and, since no special apparatus to regulate market power is required, opportunities for expropriation are limited. In these circumstances one may expect private ownership to be superior to public ownership in terms of economic efficiency, and indeed that is what the empirical evidence shows.

For a firm with market power, however, it may be desirable for reasons of allocative efficiency, and inevitable for political reasons, for privatization to be accompanied by monopoly regulation. But regulation risks blunting the very incentives – for example, for cost reduction and efficient investment – that privatization is usually intended to sharpen.

A complementary approach to the problem of privatized (and in principle also nationalized) market power is liberalization – the removal of legal and other barriers to competition, and accompanying measures to contain anti-competitive behaviour by the incumbent firm. Among other things, liberalization may expose and undermine patterns of cross-subsidy practised under public monopoly.

Therefore, in contrast to the competitive market case, it would appear that no general claim can be made as to the economic desirability of privatizing firms with market power. The accompanying regimes of regulation and competition policy are crucial determinants of the consequences of their privatization.

Privatization and Public Finance

In addition to microeconomic efficiency, considerations of public finance have motivated privatization policies in a number of countries, including Britain. By raising government revenue, privatization reduces the immediate need for public sector borrowing. It may also release firms from financial constraints resulting from government macroeconomic policy commitments. But the economic, as distinct from public accounting, significance of these points is unclear.

Selling public firms indeed raises government revenue, but the same is true of selling government bonds: in both cases the public sector receives a lump sum in return for a stream of future profit or interest payments. The deeper question is how privatization differs from government bond issue in terms of its effect on the net worth of the public sector.

If privatization leads to economic efficiency gains (which would not otherwise have been achieved) – or to greater exercise of monopoly power, which is akin to a tax increase in public finance terms – then the firm's profits are greater with privatization than in the public sector. If the firm is sold at a fair price, then the public sector captures the net present value of the profit gain (less the transactions costs of privatization, which are likely to exceed those of bonds). If, however, the firm is underpriced, then any gain to the net worth of the public sector is reduced by the extent of underpricing. Competition among potential buyers and a pre-existing market for the firm's shares are factors likely to assist more accurate pricing of privatization share issues.

Privatization can also affect the net worth of the public sector, compared with selling government bonds, if risk-adjusted discount rates differ. For example, a government with poor inflationary credibility may have to cede a large interest rate premium when selling bonds. Shares in privatized firms are not so vulnerable to expropriation via inflation (neither are index-linked bonds). However, as discussed above in relation to regulatory credibility, some privatized firms, especially those with monopoly power, may face serious risks of expropriation via regulation or even renationalization. The relative sizes of these risks of default on debt and of 'default on equity' are likely to vary by industry as well as by country. The nature of the private shareholders – for example, their nationality or whether they are small individual investors – might also be an influence upon the probability of expropriation.

Self-imposed public finance constraints by government can provide efficiency rationales for privatization if they prevent publicly owned firms from making desirable investments. In macroeconomic terms, it ought to matter little whether a firm is in public or private ownership when it does a given amount of borrowing; it appears, however, that governments seeking to adhere to public borrowing commitments may view matters differently.

Privatization and Distribution

Privatization, and the financial and industrial policies that accompany it, can have large distributional consequences. First, if public firms are sold to private investors for less than their market value – for example as part of a plan to promote 'wider share ownership' – then, relative to the situation with more accurate pricing, wealth is redistributed away from the general taxpayer to the investors who succeed in getting shares. Employees and managers of privatized firms gain from such redistribution if, as has often happened, they are allocated shares on favourable terms. Managers may benefit also from share option schemes and from being released from public sector pay constraints.

Second, if privatization hardens the firm's budget constraint, then it may diminish rents enjoyed by those within the firm to the benefit of the general taxpayer. Third, widespread cross-subsidy – for example, of small customers by large customers, and/or of suppliers of certain inputs – is a common feature of publicly owned monopoly. Privatization entails redistribution in so far as it undoes such cross-subsidies, but, here as elsewhere, the accompanying regime of regulation and competition is likely to be more important. Thus liberalization tends to be a more potent enemy of cross-subsidy than privatization itself, and, in the case of privatized monopoly, regulation can be a major determinant of the extent of redistribution among consumer groups as well as between consumers and shareholders.

Finally, it has been suggested (see Biais and Perotti 2002) that privatization policies may be designed in part so that their distributional consequences alter political preferences – in particular by giving voters a stake in the avoidance of political parties whose policies would undermine the value of shares in privatized firms.

Privatization in Practice

Privatization Worldwide

Principally since the mid-1980s, privatization policies have been pursued, to varying degrees, around the world – for example in Argentina, Brazil, Chile, France, Germany, Jamaica, Japan, Malaysia, Mexico, the Philippines, Singapore, Spain, the formerly Communist countries of central and eastern Europe and the former Soviet Union. Privatization sales proceeds worldwide are estimated to have exceeded a trillion dollars. The extensive survey by Megginson and Netter (2001) concludes that the state-owned enterprise share of global output fell from more than ten per cent in 1979 to below six per cent by 2000.

The following account concentrates on Britain, which was a leader of the worldwide privatization movement in terms of both the scale of its programme and its embrace of monopoly industries. Further details may be found in Vickers and Yarrow (1988) and Newbery (2000).

Privatization in Britain

Nationalization by the post-war Labour government and subsequently had led to a situation in 1979 where the public sector in Britain dominated the supply of energy (gas, electricity, coal and some oil), transport (air, rail and bus), communications (post and telecommunications) and water, and also had substantial interests in manufacturing (for example, in aircraft, shipbuilding, steel and cars).

In the 18 years of Conservative government from 1979 to 1997, the proportion of GDP accounted for by state-owned firms fell from 11 per cent to below two per cent. At the peak of the privatization programme, between the mid-1980s and the early 1990s, sales proceeds typically exceeded one per cent of GDP and were sometimes of the order of three per cent of public expenditure.

The watershed in the British privatization programme was the sale of British Telecom (BT) in 1984, an event motivated in good part by a desire to free BT from macro-economic policy restrictions on public sector borrowing. Before that, privatization policies were relatively modest in scale and confined to firms in more or less competitive industries such as oil and manufacturing. By extending the programme to utility monopolies, the sale of BT marked a key shift in the nature, as well as the scale, of the British privatization programme. In particular, it required the development of a system for regulating private monopoly.

Privatization – with accompanying regulation – was subsequently extended to gas (1986), airports (1987), water in England and Wales (1989), electricity (1990–91) and the railways (1996). By 1997, when the Labour Party returned to power (having abandoned its traditional commitment to public ownership), the main activities remaining in the public sector were the Royal Mail, the BBC, London Underground, British Nuclear Fuels, Air Traffic Control, and the water industry in Scotland. In 2001 National Air Traffic Services was partly privatized as a public–private partnership (PPP). London Underground remains in public ownership but since 2003 infrastructure renewal and maintenance has been procured under long-term PPP contracts.

Methods of Sale

The main ways of privatizing a firm are (*a*) offer for sale of shares to the general public, (*b*) sale to another firm, and (*c*) management/employee buyout. The third method was used in parts of the transport sector, including road haulage, some bus companies, and rail rolling stock leasing companies. The Rover car group was an example of privatization by sale to another firm (British Aerospace, which later sold Rover to BMW). However, by far the most important method used in Britain was offer for sale to the general public.

With this method, questions include (*a*) whether to sell the firm in two or more stages, or all at once; (*b*) whether the share price is set administratively or by competitive tendering among prospective purchasers of the shares; and (*c*) whether incentives and bonus schemes are created to encourage small investors to buy (and hold) privatization shares. Before the BT sale, privatizations were mostly in stages (as was BT's), use was often made of competitive tendering, and no great inducements to wider share ownership were given. These methods are conducive to reasonably accurate share pricing. Thus selling a firm's shares in stages enables accurate pricing after the first stage because the market value of the shares is known.

In the latter part of the 1980s, however, some large firms (such as British Gas) were sold in one go, tendering methods were eschewed, and there were strong incentives for small investors to buy shares. This pattern suggests that wider share ownership was a primary objective of privatization policy. In the 1990s tendering methods came back into use, albeit with discounts for small investors, thus combining the objectives of revenue maximization and wider share ownership to some extent. However, Railtrack, the railway infrastructure company, was floated on the stock market in one go in 1996.

Even judged relative to the discounts that are typical with private initial public offerings, the government revenue forgone in pursuit of the objective of wider share ownership appears to have been very large. The number of British individuals directly owning shares rose sharply but the proportion of the stock market owned directly by individuals has continued its long-run decline. If it is thought to be an appropriate policy goal, wider share ownership might be better pursued by reforms to the taxation of saving and investment generally rather than by privatization policies.

Regulation

The regulatory framework for the privatized BT was established by the Telecommunications Act 1984. A similar framework was subsequently adopted for gas, electricity, water and railways. Regulatory powers and duties were divided between the government minister, who granted licences containing regulatory provisions; an industry-specific regulator (for example, the Director General for Telecommunications), who enforced and reviewed licence conditions; and the Monopolies and Mergers Commission, which considered disputes about licence modification.

This regulatory model developed over time. Powers were transferred from individual directors general to boards, and some regulatory bodies were combined. Thus the Utilities Act 2000 created the Gas and Electricity Markets Authority, and under the Communications Act 2003 a new body, Ofcom, took over the roles of several regulators including the Director General for Telecommunications. The regulators gained powers under new UK competition law (see below). And the wider European context grew in importance, with EC directives for the liberalization of network industries such as telecommunications and energy.

For firms with market power, perhaps the most important aspect of regulation concerns price control. When it embarked on the privatization of BT the British government was anxious to avoid perceived deficiencies of rate-of-return regulation. Instead, following the report of Professor Stephen Littlechild (1983), it adopted the form of price cap regulation known as 'RPI minus *X*', which requires an index of the firm's regulated prices to fall by *X* per cent per annum in real terms (that is, relative to the retail price index) for a period of years. This was intended to be 'regulation with a light hand' and to wither away over time. However, price regulation in several industries at first became tighter and more detailed, and rate-of-return considerations were soon seen to be of prime importance at points of regulatory review. Nevertheless, even if RPI minus *X* price cap regulation is akin to rate-of-return regulation with long lags, this may well have substantial advantages over rate-of-return regulation as traditionally practised. After a time, as competition took hold after liberalization, some price controls were lifted, notably from domestic energy retail prices in 2002 (though transmission and distribution remain regulated) and from BT's retail prices in 2006.

Industry Restructuring

Restructuring is an important instrument of competition policy when firms with market power are privatized. (Forced restructuring after privatization may seriously jeopardize regulatory credibility.) Both BT and British Gas were privatized without restructuring as vertically integrated firms with nationwide dominance. However, after a decade of competition problems arising from the vertical integration of British Gas, and, in view of accelerated liberalization of retail supply, the company divided itself into separate pipeline and supply companies in 1997.

By contrast, the government radically restructured the electricity and railway industries before privatization. In 1990 the Central

Electricity Generating Board in England and Wales was divided into a transmission company (National Grid) and three generators (National Power and PowerGen; and Nuclear Electric, which was eventually privatized as British Energy in 1996). Vertical separation meant that a new mechanism had to be devised to coordinate transmission and generation, and a wholesale auction market, the Pool, was established (and later reformed by the introduction of New Electricity Trading Arrangements in 2001). In all, 12 Regional Electricity Companies (RECs) were privatized with responsibility for distribution and retail supply, which was progressively liberalized in the late 1990s and finally deregulated in 2002.

Major restructuring and ownership changes have occurred in the energy sector since privatization. The generators, National Power and PowerGen, had to divest substantial generation capacity following concerns about their market power, which was largely due to the concentrated structure for generation chosen by government in an unsuccessful effort to privatize nuclear power at the outset. Initially National Grid was jointly owned by the RECs but it became an independent company in 1995, and in 2002 merged with the gas pipeline company. After the lifting of takeover protections in the mid-1990s, most RECs were acquired, and ten years later six companies, supplying both gas and electricity (often in combined deals), accounted for nearly all energy supply – British Gas and five electricity suppliers, of which one is French- and two are German-owned. Thus, depending on merger policy, industry structure and ownership can alter substantially after privatization.

British Rail was restructured before privatization to separate network infrastructure from train operation. Railtrack, which took over network infrastructure, including track and stations, was privatized in 1996. The company went into administration in 2001 and its assets were acquired by Network Rail, a company limited by guarantee that has no shareholders. Three rolling stock leasing companies were also privatized in 1996 (and soon resold at a profit). Private train operating companies run train services under franchises. Large public subsidy to rail services continues in the privatized regime.

Liberalization of Competition

Statutory monopoly typically accompanied public ownership in the utility industries. Among other things this served to facilitate extensive cross-subsidy between groups of customers, and sometimes of input suppliers – for example, the nationalized electricity industry effectively subsidized British Coal. The removal of statutory barriers to entry in telecommunications, gas and electricity began in the early 1980s, before privatization policies were adopted, but then had little competitive effect. Liberalization has generally gone further since privatization – as illustrated above by the energy sector – and over time more attention has been given to economic, as well as legal, barriers to entry.

In telecommunications, liberalization of apparatus supply and value-added services began in 1981, when BT was split from the Post Office, and in 1982 Mercury was licensed as a competing network operator. However, for the rest of the decade the government adopted a ‘duopoly policy’ of allowing no further entry into fixed-link network operation. A parallel duopoly policy applied to mobile telecommunications.

When the duopoly policy was ended in 1991, the interconnection question – on what terms can rivals gain access to BT’s local network? – became and has remained a focus of controversy. On the one hand it was argued that rivals could inefficiently ‘cream-skim’ BT’s more profitable business while BT remained restricted by controls on its tariff structure and universal service obligations. On the other hand, it was argued that rivals faced entry barriers. These tensions eased somewhat over time as tariff rebalancing diminished cross-subsidies in BT’s pricing structure, and as entry barriers (such as the lack of number portability) were tackled directly by the regulator. But the advent of broadband, with BT still an integrated incumbent operator, brought the interconnection question back into sharp focus. Faced with the prospect of an investigation under competition law, BT agreed in 2005, 20 years after privatization, to operational separation of its local access infrastructure.

A major weakness of UK policy towards privatized firms with market power had been the absence of effective competition law against anti-competitive agreements and abuse of dominance. However, that gap was filled in March 2000 when the Competition Act 1998 – which mirrors Articles 81 and 82 of the EC Treaty – came into force and was followed by the Enterprise Act 2002. The regulators can now apply (non-merger) competition law in their sectors. Over time, then, following the shift from state monopoly to regulated private monopoly, there has been increasing availability of competition policy instruments to address market power in historically monopolized industries such as energy and telecommunications in Britain. Nevertheless, the regulatory regimes have remained the principal means of controlling market power.

The Performance of Privatization

Privatization policies have undoubtedly had major economic and financial effects. Have they generally been positive? Answering this question properly requires the specification of evaluation criteria, performance measures, statistical methods and the counterfactual: what would have happened without privatization?

Megginson and Netter (2001, section 5) review 38 empirical studies of privatization covering both developed market economies and transition economies. Privatized firms are generally found to become more efficient and profitable, and to invest more. There are mixed results on employment effects, though job cuts appear to be associated with corresponding productivity gains. Direct evidence on effects on consumers is limited. In their survey of studies of transition economies, Djankov and Murrell (2002) conclude that privatization, especially to outside investors as distinct from managers and workers, is robustly associated with enterprise restructuring and growth, and that competition has a significant positive effect on enterprise performance.

In competitive industries, improvements in the corporate performance of privatized firms imply overall economic gain, and there is ample evidence that privatization has been a success. For firms with market power, however, corporate performance can improve at the expense of the public as well as by enhanced efficiency. Moreover, it is hard to isolate the effects of privatization in hitherto monopolized industries from those of accompanying regulatory and competitive reforms. In Britain, methods of privatization and regulatory reform have at times been seriously flawed. But privatization was probably necessary for liberalization and for the creation of a system of independent economic regulation, augmented in time by effective competition policy. Though far from perfect, these are major improvements upon the nationalized monopoly of old.

See Also

- ▶ Competition
- ▶ Policy Reform, Political Economy of
- ▶ Public Utility Pricing and Finance
- ▶ Regulation

Bibliography

Biais, B., and E. Perotti. 2002. Machiavellian privatization. *American Economic Review* 92: 40–258.

Djankov, S., and P. Murrell. 2002. Enterprise restructuring in transition: A quantitative survey. *Journal of Economic Literature* 40: 739–792.

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

Littlechild, S.C. 1983. *Regulation of British telecommunications profitability*. London: HMSO.

Megginson, W.L., and J.M. Netter. 2001. From state to market: A survey of empirical studies on privatization. *Journal of Economic Literature* 39: 321–389.

Newbery, D. 2000. *Privatization, restructuring and regulation of network utilities*. Cambridge, MA: MIT Press.

Schmidt, K.M. 1996. The costs and benefits of privatization: An incomplete contracts approach. *Journal of Law, Economics, and Organization* 12: 1–24.

Vickers, J., and G. Yarrow. 1988. *Privatization: An economic analysis*. Cambridge, MA: MIT Press.

Privatization Impacts in Transition Economies

Saul Estrin

Abstract

This article addresses the large-scale privatization processes in central and eastern Europe. It explains why reformers placed such emphasis on privatization and the practical problems posed by the scale of state ownership under communism, leading to the widespread use of mass privatization. As a result ownership changes were huge and extremely rapid but the improvement in corporate governance was more questionable. The empirical findings about the impact on enterprise performance are patchy, though on balance the effect has been positive, especially in countries with stronger institutions or where the new owners have been foreigners.

Keywords

Agency chains; Asymmetric information; Capital market; Company performance; Concentration of ownership; Corporate governance; Corruption; Foreign direct investment; Free rider problem; Growth; Insider ownership; Intermediate agents; Management–employee buyouts; Monitor; Opportunity costs; Outsider ownership; Ownership and control; Privatization; Privatization in transition economies; Privatization vouchers; Productive efficiency; Property rights; Rent seeking; Restitution; Soft budget constraint; Takeovers; Transition economies

JEL Classifications

P3

Privatization is the process whereby the ownership of the state's productive assets, often utilities or large industrial enterprises, is transferred into private hands. This has been a major activity for governments in both the developed and the developing worlds since Prime Minister Thatcher's first modern privatization programme in the UK between 1979 and 1984. The cumulative revenues raised from the process globally probably exceeds $1.25 trillion dollars, while the role of state-owned enterprises in the economies of high income countries has declined from around 8.5% GDP on average in 1984 to around 6% in 1991 and probably below 5% in 2005 (see Megginson 2005). The reduction in state ownership has probably been even more dramatic in less developed countries, from around 16% GDP in 1981 to around 5% in 2004. Privatization is intended to improve corporate efficiency and generate revenues for the state, and there is now probably sufficient experience in different economic and institutional environments to evaluate its impact relative to expectations.

Privatization has been a particularly important phenomenon in the transition process in central and eastern Europe from planning to a market system. This is because Communist regimes had placed almost all the productive assets of the economy in state hands for ideological reasons, and to facilitate the planning process. As a result, countries like Czechoslovakia and the Soviet Union contained virtually no private sector at all – typically in excess of 90% of assets were state owned and even in countries with slightly larger private sectors, like Poland or Hungary, private ownership was concentrated in agricultural and handicraft activities; industrial firms were all in state hands. This meant that privatization was a central aspect of building a market economy in all the transition economies. Indeed, to quote Dusan Triska in 1992, 'privatization is not just one of the many items on the economic program. It is the transformation itself' (see Estrin 2002).

This article addresses the privatization process in central and eastern Europe, focusing on the objectives, the methods and, most importantly, the impact of the ownership changes. Privatization always had some ideological content in the transition economies, especially in the early years, when the reformers wished to create a 'capitalist class' supporting the radical changes that were required to build a market economy. But the fundamental objective of privatization in transition economies, as in developed and developing ones, has been to enhance company performance. We enquire whether privatization has succeeded in this objective in the remainder of this article.

Why Privatize?

We begin by identifying why reformers in the transition economies placed such emphasis on privatization. Transforming state-owned assets into private hands can improve corporate efficiency (see Vickers and Yarrow 1985), and, particularly with the privatization of infrastructure, the benefits can spill over to the rest of the economy. To understand why, one must compare company objectives and corporate governance under state and private ownership. It is normally argued that the fundamental difference between state-owned and private firms rests in their objectives: the latter focus exclusively on profit, which generates close attention to costs and to the demands of customers. State-owned firms may be interested in profits too, but they will almost certainly be expected by their owners to satisfy other objectives as well, for example, politically determined targets such as creating or maintaining employment in economically depressed regions or holding prices below average costs for redistributive reasons. In this situation, profits become a secondary criterion, or indeed an irrelevance, and business decisions become politicized. Inefficiencies can thrive because they are not a central concern of the owner, and managers can exploit the lack of clarity in company objectives to ensure an easy life for themselves and employees (see Shleifer and Vishny 1994).

Therefore, an important motive for privatization is to focus attention on profits as the sole objective for the enterprise sector. But the problems of state ownership go beyond just diffuse and non-commercial objectives. In a socialist economy, the system of administered prices also means that privatization and market liberalization are needed to reveal opportunity costs. Moreover, even in a market economy, when a public-sector firm operates in a competitive market and the government tries to enforce an objective of profit maximization on its management, weaknesses in corporate governance can still cause inferior performance to what might be achieved under private ownership. The problem is centred on the asymmetry of information held by managers and owners; outside owners – private or state – can never have full access to the information about corporate performance that is in the hands of managers. Thus, it is hard for them to establish whether poor results are a consequence of unforeseen circumstances or managers exploiting firm profits for their own purposes. Whenever ownership and control are separated, firm-specific rents can be used to satisfy management's aim – for example, lower effort or managerial power, via the size of the firm – rather than profits. However, a private ownership system places more effective limits than does state ownership on their discretionary behaviour, via external constraints from product and capital markets which largely operate through the market for corporate control, and through the internal constraints imposed via statutes and monitoring by the owners themselves (see Estrin and Perotin 1991).

In Anglo-Saxon countries, the constraints on managerial discretion in large part derive from stock markets (see Megginson 2005). The quality of managerial decision-making and the extent of managerial discretion are an input in the choices of traders in equity markets, whose judgement on company performance is summarized in the share price. If the managerial team is thought to be incompetent or inefficient, the share prices will be reduced, putting pressure on managers to improve their performance. A persistently poor showing by a quoted company may also generate external pressure by encouraging a takeover bid. In this case, the stock market can be viewed as a market for corporate control, with alternative teams vying for the right to manage the enterprise. However, the effectiveness of these disciplines relies to some extent on the concentration of ownership. If ownership is highly dispersed, each individual owner has only a slight incentive to monitor effectively, and as there is a free rider problem monitoring may be inadequate.

Governance also comes from the way that the managerial market operates, with managerial performance, pay and job prospects assessed by movements in share prices. Payment mechanisms such as management stock option schemes can also be put in place to align the incentives of owners and managers. In countries such as Japan or Germany, however, the mechanisms can be different, with less reliance on an adversarial market for corporate control and more extensive use of internal governance constraints. Ownership is typically highly concentrated in the hands of banks, funds or families who are granted board representation and undertake close monitoring of managerial performance directly, and use the managerial market and management incentive schemes.

Either way, it is hard for the state to imitate these market-based constraints. State-owned firms are not subject to private capital market disciplines, so neither the competitively driven informational structure nor the market-based governance mechanisms can be substituted for in full. State employees are usually civil servants and do not compete in the wider managerial market, though Western governments have recently tried to reduce the labour market segmentation between the public and private sectors. Moreover, though the government's ownership stake is concentrated, the state is rarely directly represented on the boards of public sector companies and usually does not have the capacity in the supervisory ministries to undertake the necessary scale and quality of monitoring (see Vickers and Yarrow 1985).

These arguments have particular resonance in the transition economies of central and eastern Europe. The economic problems of the socialist system were largely a result of the impact of state ownership and planning on investment allocation, incentives and efficiency (see Gregory and Stuart 2004). Firms did not attempt to maximize profits, and productive efficiency was a low priority. Instead, weak monitoring of managers by the state as owner and the absence of external constraints gave management almost total discretion to follow their own objectives – rent absorption, asset stripping, employment, social targets. The softness of budget constraints (Kornai 1990) that goes with the political determination of resource allocation was a further source of incentive problems, since managers did not have to bear the consequences of their own actions. Mistakes were condoned and losses were subsidized.

Methods of Privatization in Transition Economies

It is therefore clear why privatization was so important in the transition process. Nonetheless, reforming governments might in principle have left privatization until the track records of particular firms in the market environment had become firmly established and until the stock of domestic savings in private hands was sufficient to ensure the success of a competitive bidding process for the assets. But the state was probably not able to manage its assets effectively in the intervening period, and managers and workers began very rapidly to steal the assets (Canning and Hare 1994). The collapse of communism had left state-owned firms with limited internal structure to handle the new requirements of the marketplace and no mechanisms to monitor or enforce governance on state-owned firms (see Blanchard et al. 1991). The authorities had either quickly to create structures whereby the state as owner could control enterprise decisions or face a gradual dissipation of the net worth of the enterprise sector by consumption, waste or theft. These stark alternatives persuaded many reforming governments and their Western advisors to consider rapid privatization. (Boycko et al. 1995 – the first of these an insider to Russia policymaking at the time – make a similar point concerning Russia. They argue that Russia had to undertake a massive and speedy ownership change in order to break the tradition of rent-seeking behaviour and the long-standing links between the state and the enterprise sector. They argue that, in order to gain political support for the privatization process, substantial stakes had to be given to insiders – the managers and workers in firms – so that they did not block the process.)

The sheer scale of privatization required in the transition economies posed considerable practical problems. As we have seen, the Communist heritage meant that the majority of firms in the economy needed to be privatized. At the aggregate level, the stock of domestic private savings in these countries was too small to purchase the assets being offered. This led the reformers to innovate with privatization methods.

For selected firms, many transition economies used auction or public tender, as have been the norm in the West. Such sales could in principle be to domestic or foreign purchasers but, in practice, only Hungary and Estonia were willing or able to sell an appreciable share of former state-owned assets to foreigners. Foreign capital ended up purchasing about 20% of the privatized assets in Hungary and up to 50% in Estonia, but even in these countries the preponderance of foreign ownership gave rise to public disquiet. Moreover, foreign direct investment flows to the transition economies were modest in the early years, when privatization was taking place, and were highly concentrated towards the Czech Republic, Hungary and Poland (United Nations 2004; Meyer 1998). In practice, sales of state-owned enterprises have mainly been to a country's own citizens: either to external capital owners or to insider management–employee buyouts. Managers and employees were the more common initial buyers, perhaps because they had insider knowledge about their company's business prospects. Some governments, such as in Romania, actively encouraged the emergence of insider-owned firms.

Some countries also experimented with restitution to former owners; the former East Germany, Hungary, the former Czechoslovakia and Bulgaria are prominent examples. Restitution has the advantage that it immediately creates a property-owning middle class and re-establishes 'real ownership'. However, the process of restitution entails legal complexities. For example, suppose that a factory has been built on a plot of land formerly owned by a noble. Does the noble receive the land back, and therefore rental for the factory? Or should the noble be compensated for the value of the property at the time of its seizure and, if so, how is such an evaluation to be made some 80 years later? Restitution also raises the deep question of how the assets accumulated during the Communist era, when consumption levels were held down for national capital accumulation, should be distributed. Since the burden of lower consumption was imposed on everyone, the argument that the distribution of the resulting assets should be egalitarian has been a powerful one.

To increase the pace of privatization, a number of transition countries began to experiment with 'mass privatization'. This entails placing into private hands nominal assets of a value sufficient to purchase the state firms to be privatized. To avoid the inflationary consequences of such wide-scale 'money' creation, the new assets must be non-transferable and not valid for any transaction other than the purchase of state assets. This was largely achieved using the instrument of privatization vouchers or certificates. It was hoped that any deficiencies in the resulting corporate governance mechanism arising from the fact that the ownership structure was initially diffuse would be addressed by capital market pressures leading to increased ownership concentration (Boycko et al. 1995).

Mass privatization has been carried out in a number of different ways, but the differences can be summarized around two issues. The first was whether the vouchers or certificates were distributed on an egalitarian basis to the population as a whole or whether, as in Russia and many other countries of the former Soviet Union, management and employee groups received many of the shares, perhaps to diffuse potential opposition to privatization. Second, policymakers needed to determine whether vouchers were intended to be exchanged directly for shares in companies, or whether the vouchers should be in funds that own a number of different companies. In the Czech and Slovak republics and in Russia, vouchers were exchanged directly for shares, although financial intermediaries soon developed in the market. In the Polish scheme, vouchers were exchanged for shares in government-created funds that jointly owned former state-owned enterprises.

Every country used a variety of privatization methods and everywhere different sorts of firms were sold in different ways. For example, in most transition economies small firms were usually sold to the highest bidder, and utilities were often floated on stock markets. However, it was possible by the time the bulk of privatization was completed in the late 1990s to discern the predominant method used in each country, and we report the most widely used summary in Table 1 from the EBRD's *Transition Repor*t, 1998. Mass privatization was the most common privatization method across the transition economies; 19 of the 25 countries listed used some form of mass privatization as either a primary or secondary method. Moreover, management–employee buyouts (MEBOs) also proved important, perhaps because transition governments sometimes did not have the authority to take on entrenched insiders in firms. Thus, nine countries used MEBOs as their primary method, and six as their secondary method. Most transition economies therefore eschewed the conventional method of privatization, by direct sale. In fact, only five countries used this as their primary privatization method, though these were among the most developed transitional economies.

The Scale of Privatization

There was an extremely speedy ownership change in most transition economies. Few countries had contained a private sector of any significance in

Privatization Impacts in Transition Economies, Table 1 Methods of privatization

	Primary method			Secondary method		
Country	Direct sales	MEBOs[a]	Vouchers	Direct sales	MEBOs[a]	Vouchers
Albania		+				+
Armenia			+		+	
Azerbaijan			+	+		
Belarus		+				+
Bulgaria	+					+
Croatia		+				+
Czech Republic			+	+		
Estonia	+					+
FYR Macedonia		+		+		
Georgia			+	+		
Hungary	+				+	
Kazakhstan			+	+		
Kyrgyzstan			+		+	
Latvia			+	+		
Lithuania			+	+		
Moldova			+	+		
Poland	+				+	
Romania		+		+		
Russia			+	+		
Slovak Republic	+					+
Slovenia		+				+
Tajikistan		+				+
Turkmenistan		+		+		
Ukraine			+		+	
Uzbekistan		+		+		

[a] Management-employee buyouts

Source: EBRD (1998)

1990. Exceptions were Hungary and Poland, where there had been long-standing private firms in agriculture and crafts, and the private sector already represented over 30% of GDP (see Estrin 1994). But in the transition economies as a whole the private sector contribution to GDP was usually less than 20%. The growth in the private sector share during the 1990s, reported in Table 2, is extraordinary. As early as 1995, the private sector share was above 50% in nine countries, though in eight former republics of the Soviet Union it remained below 30%. By 2002, the private sector in 13 additional nations had reached at least 50% of GDP and in only two laggards, Belarus and Turkmenistan, was private sector activity still below 25% of GDP. Thus the privatization process in the transition economies was in many countries effective in transferring the bulk of economic activity from state to private hands in the space of hardly more than a decade.

This remarkable performance should not conceal real concerns raised at the time about the quality of privatization, and therefore about its consequences for enterprise restructuring. First, there are questions about how real the privatization has been. In many transition economies, the state continued to own golden shares or significant shareholdings in companies. For example, the Russian state retained more than a 20% share in 37% of privatized firms, and kept more than a

Privatization Impacts in Transition Economies, Table 2 Private sector percentage shares in GDP and employment, 1991–2002

	In GDP			In employment		
	1991	1995	2002	1991	1995	2001
Albania	24	60	75	–	74	82
Armenia	–	45	70	29	49	–
Azerbaijan	–	25	60	–	43	–
Belarus	7	15	25	2	7	–
Bosnia and Herzegovina	–	–	45	–	–	–
Bulgaria	17	50	75	10	41	81
Croatia	25	40	60	22	48	–
Czech Republic	17	70	80	19	57	70
Estonia	18	65	80	11	–	–
Fyr Macedonia	–	40	60	–	–	–
Georgia	27	30	65	25	–	–
Hungary	33	60	80	–	71	–
Kazakhstan	12	25	65	5	–	75
Kyrgyz Republic	–	40	65	–	69	79
Latvia	–	55	70	12	60	73
Lithuania	15	65	75	16	–	–
Moldova	–	30	50	36	–	–
Poland	45	60	75	51	61	72
Romania	24	45	65	34	51	75
Russia	10	55	70	5	–	–
Serbia and Montenegro	–	–	45	–	–	–
Slovak Republic	–	60	80	13	60	75
Slovenia	16	50	65	18	48	–
Tajikistan	–	25	50	–	53	63
Turkmenistan	–	15	25	–	–	–
Ukraine	8	45	65	–	–	–
Uzbekistan	–	30	45	–	–	–
Means	20	44	62			

Source: EBRD (1999, 2003)

Privatization Impacts in Transition Economies, Table 3 Percentage of privatized firms with retained state shareholdings

	Percentage of shares retained by the state		
Country	0%	1–30%	> 30%
Albania	83.9	0	16.2
Armenia	97.1	2.9	0
Azerbaijan	94.1	5.9	0
Belarus	80.4	10.7	8.9
Bulgaria	30.8	61.6	7.7
Croatia	59.1	33.4	7.6
Czech Republic	100	0	0
Estonia	92.3	3.9	3.9
Georgia	79.3	7	13.8
Hungary	100	0	0
Kazakhstan	93.6	2.1	4.2
Kyrgyz Republic	91.2	1.8	7.1
Latvia	100	0	0
Lithuania	80.8	19.3	0
Macedonia (FYR)	92.9	0	7.1
Moldova	87.7	5.4	7.1
Poland	71.7	13.3	15.1
Romania	80	13.4	6.7
Russia	82.6	8.7	8.7
Slovak republic	92.3	7.7	0
Slovenia	63	24.1	13
Ukraine	83.6	9.6	6.8
Uzbekistan	69.2	27	3.9
Total	80.9	11.8	6.3

Source: Unpublished EBRD survey, used by Bennett et al. (2007)

40% share in 14% of the firms that it privatized. Only in half of privatized firms did the Russian government sell its entire holding. Thus, the clean break between the state as owner and the enterprise sector has perhaps been more notional than real. In a survey of privatized firms undertaken by the EBRD in 1999, reported in Table 3, we show that in 20 of the 23 countries the state has retained some shares post-privatization. On average, the state retained some shares in around 20% of privatized firms, with more than a 20% shareholding in around 12% of the firms. It is suggestive that retained state shareholdings are negligible in some of the leading transition economies – for example, the Czech Republic, Hungary and Latvia – but the state has tended to keep a larger share in less advanced transition economies: more than 15% of privatized firms in Albania, Belarus, Georgia, Lithuania, Poland, Romania, Russia and Ukraine, and more than 30% of privatized firms in Bulgaria, Croatia, Slovenia and Uzbekistan. State ownership has also been retained in many developed OECD economies, including via the use of 'golden shares'. According to Bortolotti and Faccio (2006), governments were actually the largest stakeholder or held special control powers (golden shares) in 62.4% of privatized OECD companies.

But widespread retained state ownership is not the only indication that privatization may not have ensured the establishment of effective corporate governance mechanisms in transition economies. The long 'agency chains' implicit in mass privatization may not provide appropriate incentives for corporate governance. Voucher privatization led to ownership structures that were highly dispersed (Coffee 1996). Typically the entire adult population of the country, or all insiders to each firm, were allocated vouchers with which to purchase the shares of the company. The desire for equitable and politically acceptable outcomes dominated the need to create concentrated external owners who would have a large enough stake to be motivated to maintain oversight of management. However, it was possible that financial intermediaries could aggregate individual voucher holdings and carry out effective monitoring of management, and in Czech Republic, Poland, Slovenia and Slovakia some effort was made to ensure such concentrated intermediate agents did emerge. This was often associated with fraud and the outright theft of assets by managers to avoid their use by the new owners – so-called 'tunnelling' (Johnson et al. 2000).

The way that mass privatization was carried out in many countries also sometimes led to majority ownership that was not best suited to accelerate restructuring, for example by insiders. This was probably largely for political reasons, especially in countries where the pro-reform forces were politically weak. According to Earle et al. (1996), insiders held a majority shareholding in 75% of firms in Russia immediately post-privatization (1994) and outsiders only 9%. Insider ownership was predominantly in the hands of workers. However, this created little problem for management because worker ownership was so highly dispersed. Indeed Blasi et al. (1997) argue that control was effectively in the hands of management in Russian employee-owned firms. Outsider ownership is also typically highly dispersed, with much of it in the hands of banks, suppliers, other firms and an assortment of investment funds. In Russia, it appears from a variety of studies (see Estrin and Wright 1999, for a survey) that outside shareholding has increased at the expense of the state and insiders during the 1990s, but ownership is also becoming increasingly dispersed and the greater degree of outside ownership may largely represent the fact that former insider voucher owners have left the firm but retained their shares.

This pattern of extensive employee ownership seems broadly consistent with the evidence for other CIS countries. In Ukraine, insiders owned 51% of shares in all privatized firms in 1997 – managers 8% and workers 43% – while outsiders held 38% and the state residue share was 11%. In Ukraine, insiders have actually increased their shareholdings, while managers have been buying shares from workers. Thus, rather than evolving towards the structure of firms owned by a concentrated group of outsiders, as was hoped by reformers, enterprises in the CIS appear to have remained primarily owned by dispersed groups of employees or outsiders. However, the situation appears to have been somewhat different in central Europe, where many of the most important firms in the economies are now quoted on the relevant national stock exchanges or owned by large foreign firms – for example, Skoda and Volkswagen. As we have seen, foreign ownership was predominant in Hungary, ownership by new entrepreneurs was common in Poland, while investment-fund ownership predominated in the Czech Republic.

The Impact of Privatization

In this section, we analyse the impact of privatization on economic and company performance in the transition economies. This can be considered from the macro-economic and the microeconomic sides, and we provide some information on both. We start by considering the effects on government resources, and exploring the relationship between private sector shares, privatization methods and revenues and economic growth. We then summarize the findings of the very large literature about the effects of privatization on company performance.

In Table 4, we present the cumulative revenues from privatization in each of the transition economies, from 1995 to 2002. The sums were relatively modest in most countries in 1995; cumulative revenues were less than 2% of GDP in 15 countries of the 23 covered, and exceeded 20% in only one country, namely, Hungary. The situation had changed appreciably by 2002. Cumulative revenues from privatization exceeded 5% of GDP in 14 countries, exceeded 10% of GDP in eight and were greater than 30% in Hungary and Slovakia. Thus, even in countries which used mass privatization the selling of state assets proved to be a significant source of government revenue through the financially demanding period of early transition, and may therefore have

Privatization Impacts in Transition Economies, Table 4 Privatization revenues (cumulative, in percentage of GDP), 1995–2002

	1995	1996	1997	1998	1999	2000	2001	2002
Albania	3.1	3.3	3.6	3.6	3.9	7.0	9.1	9.1
Armenia	3.4	3.4	3.4	5.6	6.7	8.8	9.4	9.7
Azerbaijan	0.0	0.1	0.3	0.9	1.5	1.7	2.0	2.4
Belarus	0.5	0.7	0.9	1.0	1.1	1.1	1.2	2.9
Bosnia and Herzegovina	0.0	0.0	0.0	0.0	0.7	2.0	2.8	2.9
Bulgaria	0.7	1.5	4.6	6.2	8.4	9.7	10.3	11.2
Croatia	0.9	1.4	2.0	3.6	8.2	10.2	13.5	15.8
Czech Republic	4.6	6.3	7.1	7.9	9.3	10.3	13.1	18.7
Estonia	0.0	0.0	0.2	0.3	4.2	5.2	7.2	7.6
Georgia	19.1	19.8	20.5	21.8	22.7	23.0	23.1	–
Hungary	20.8	23.4	27.5	28.6	29.8	30.2	30.6	30.6
Kazakhstan	3.7	5.9	9.2	13.0	14.8	15.6	16.1	16.6
Kyrgyz Republic	0.9	1.3	1.4	1.6	1.9	2.1	2.5	2.7
Latvia	0.7	0.8	2.2	3.3	3.5	4.1	4.7	5.4
Lithuania	1.4	1.4	1.6	6.8	8.0	9.8	10.8	11.3
Moldova	0.8	1.3	3.6	4.4	5.4	11.1	11.1	–
Poland	2.6	3.6	5.1	6.4	7.7	11.4	12.2	12.6
Romania	1.2	2.2	4.6	6.4	7.6	8.2	8.9	9.0
Russia	1.5	1.7	2.7	3.4	3.5	3.8	4.2	4.5
Slovak Republic	8.4	10.2	10.8	11.5	11.8	16.3	20.1	35.1
Slovenia	0.4	0.9	1.4	2.2	2.5	2.5	2.7	4.9
Tajikistan	1.5	1.7	2.3	2.8	3.6	4.6	4.8	5.8
Turkmenistan	0.2	0.2	0.2	0.2	0.3	0.6	0.6	0.6

Source: EBRD (2004)

contributed to macro-stability and growth. However, there is no empirical evidence linking growth to the private sector share. Bennett et al. (2007) explore the impact of privatization on growth, but, while they identify a positive effect from the use of the mass privatization method, they do not find any significant relationship between growth and the private sector share. There is a limited amount of academic work for other economies, which explores the impact of privatization on growth rates. In an early study, Plane (1997) looks at the effects of divestiture on growth in a sample of 35 developing countries. He also controls for the problem of reverse causality by identifying separately the factors that determine a successful privatization programme. He finds that the impact of privatization on economic growth is indeed positive, and is strengthened when privatization occurs in infrastructure or in industrial sectors. Zinnes et al. (2001) use a fairly short sample period to undertake an aggregate growth study for the transition economies. They conclude that, while privatization does not actually increase growth, there is a positive impact when the privatization process is accompanied by institutional reforms.

To turn to the microeconomic evidence, there have been a large number of studies of how privatization affects the performance of firms in transition economies. The most complete of these is by Djankov and Murrell (2002), which surveys the findings of more than 100 empirical studies of transition economies and uses a meta-analysis of the results to draw conclusions. Despite the plethora of material, the overall findings remain ambiguous. This is partly because the studies employ a variety of data-sets, measurements and methods which produce contradictory results. For example, there are many ways of measuring company performance, including profitability, productivity, sales growth, export growth, and restructuring; and their findings differ. To begin with productivity, there is a wide variance in results across countries and samples, with private ownership found to yield positive, zero or negative effects. There is, however, convincing evidence that sale to foreign owners yields a positive effect, and that privatization is more likely to improve performance in central Europe than in the former Soviet Union. More recent literature strongly confirms the results with respect to foreign direct investment (for example, Sabirianova et al. 2005). However, it is harder to discern positive effects from privatization when profitability is the performance measure, though once again some studies find a positive impact when the firm is sold to a foreign owner, and very few studies isolate a positive significant effect of privatization on revenues. There are fewer studies of the impact of privatization on exports, and these tend to be positive, especially when foreigners take over the former state-owned firm, and restructuring activity seems to have been significant in privatized firms in central Europe, but not in Russia, Ukraine and other countries of the former Soviet Union.

The variation in results is not merely a consequence of the wide variety of measures and countries with which the effects of privatization have been tested. Some serious methodological problems bedevil work of this sort, most importantly that of selection. This is the situation when firms with particular characteristics – for example, superior performance – were systematically chosen for privatization. In such a case, while one observes what appears to be superior performance among firms that have been privatized, the correct interpretation is not that privatization enhances performance but that it was the better firms that were chosen for privatization. The converse applies if the state chooses to keep the best firms for itself and to sell only the less productive ones; in this case, privatization will appear to lead to worse performance. Unfortunately, very few studies of privatization in the transition economies have been able to do much to address this problem of reverse causality. The data-sets upon which the empirical work has been based have been small and usually derived from sample survey questionnaires that did not contain sufficient information to control for the selection problem.

Even so, Djankov and Murrell (2002) conclude on the basis of the weight of the evidence that the impact of privatization on company performance has probably been positive and significant, though not in every circumstance. Two factors are usually cited as being particularly influential in determining whether privatization acts to enhance company performance. The first is the nature and characteristics of the new private owners. We noted that foreign owners lead to an improvement on most measures of performance. There is also some evidence, though it is less convincing, that sale to domestic private owners also improves performance, though it can be important for the ownership shares to be concentrated. However, there is almost no evidence that company performance is improved when firms are sold to insiders, either managers or workers. This is probably because insiders have exploited their control to resist the changes in behaviour required to make firms competitive in the market environment, rather than to promote them. We observed above that insider ownership was a fairly common phenomenon, especially in the former Soviet Union, and this probably goes some way to explain why economic performance in many of those countries was weaker than in, for example, much of central Europe, such as Hungary, Poland and the Czech Republic, where foreign direct investment flows were much greater.

The second factor is the institutional and business environment in which privatization takes place. We noted above that privatization relies on improved corporate governance, but that in turn depends on a competitive market environment and the enforcement of property rights. In countries where the legal system is not functioning effectively, and businesses face high level of corruption and weak standards of financial discipline, it is hard to imagine how private ownership on its own might be expected to improve company performance. For example, sharper performance is meant to come from tighter financial disciplines to eradicate waste and reduce cost, but these will not bind in situations when budget constraints remain soft, as occurred post-privatization in many countries of the former Soviet Union, with firms financing their deficits not through direct government subsidies but by not paying their bills, especially to their workers, to the government in taxes, and to the state-owned utility companies.

These two limiting factors affect some privatizations in all transition economies, but on average have been more likely to pertain in the economies of the former Soviet Union than those of central and eastern Europe. Thus while the macroeconomic work suggests a clear positive impact from privatization on economic growth, the results from the microeconomic literature are more modulated. The positive effects from privatization are found not to be automatic. They depend on to whom the firm was sold – foreigners, outsiders or insiders – and on the broader business environment in which the firm operates. The latter in particular tends to be better in central Europe and especially in the new accession economies to the European Union. Privatization methods may also have played an important role (see Bennett et al. 2007).

Conclusion

The most impressive feature of privatization in the transition economies has been the speed and scale at which it occurred. The reforming governments of the late 1980s and early 1990s managed successfully to transfer the huge state-owned sector into largely private hands in a time period of hardly more than a decade, and to do so they had to use innovative privatization methods. However, this led them to introduce private ownership into situations where other crucial aspects of the business environment were not yet sufficiently developed to support the private economy. We find that privatization appears to have provided governments with much-needed revenues. However, at the enterprise level the results on performance are more patchy, though on balance the effects of privatization have probably been positive, especially when the new owners were foreigners. The most serious problem for privatization as a policy has been its use in a weak legal and institutional environment. In such cases, it rarely appears to have improved company performance.

See Also

- ▶ Corporate Governance
- ▶ Privatization
- ▶ Transition and Institutions

Bibliography

Bennett, J., S. Estrin, and G. Urga. 2007. Privatization and economic growth in transition economies. *Economics of Transition* 15: 661–683.

Bevan, A., and S. Estrin. 2004. The determinants of foreign direct investment into European transition economies. *Journal of Comparative Economics* 32: 775–787.

Blanchard, O., R. Dornbusch, P. Krugman, R. Layard, and L.H. Summers. 1991. *Reform in Eastern Europe*. Cambridge, MA: MIT Press.

Blasi, J., M. Kroumova, and D. Kruse. 1997. *Kremlin capitalism: Privatizing the Russian economy*. Ithaca: Cornell University Press.

Bortolotti, B., and M. Faccio. 2006. Reluctant privatization. Working paper no. 40/ 2004, ECGI.

Boycko, M., A. Shleifer, and R. Vishny. 1995. *Privatizing Russia*. Cambridge, MA: MIT Press.

Canning, A., and P. Hare. 1994. Hungary. In Estrin (1994).

Coffee, J. 1996. Institutional investors in transitional economies: Lessons from the Czech experience. In *Corporate governance in Central Europe and Russia*, ed. R. Frydman, C. Gray, and A. Rapaczynski, vol. 1. Budapest: Central European University Press.

Djankov, S., and P. Murrell. 2002. Enterprise restructuring in transition: A qualitative survey. *Journal of Economic Literature* 40: 739–793.

Earle, J., S. Estrin, and L. Leschenko. 1996. Ownership structures, patterns of control and enterprise behaviour in Russia. In *Enterprise restructuring and economic policy in Russia*, ed. S. Commander, Q. Fan, and M. Schaffer. Washington, DC: Economic Development Institute and World Bank.

EBRD (European Bank for Reconstruction and Development). 1998. Various years. *EBRD transition report*. London: EBRD.

Estrin, S., eds. 1994. *Privatization in Central and Eastern Europe*. London: Longman.

Estrin, S. 2002. Competition and corporate governance in transition. *Journal of Economic Perspectives* 16(1): 101–124.

Estrin, S., and V. Perotin. 1991. Does ownership always matter? *International Journal of Industrial Organization* 9: 55–72.

Estrin, S., and M. Wright. 1999. Corporate governance in the former Soviet Union: An overview. *Journal of Comparative Economics* 27: 398–421.

Gregory, P.R., and R.C. Stuart. 2004. *Comparing economic systems in the twenty-first century*. New York: Houghton Mifflin.

Johnson, S., R. La Porta, F. Lopez-de-Silanes, and A. Shleifer. 2000. Tunneling. *American Economic Review* 90: 22–27.

Kornai, J. 1990. *The road to a free economy. Shifting from a socialist system: The example of Hungary*. New York/ Budapest: Norton/HVG Kiadó.

Megginson, W.L. 2005. *The financial economics of privatization*, 1st ed. New York: Oxford University Press.

Meyer, K. 1998. *Direct investment in economies in transition*. Aldershot: Edward Elgar.

Plane, P. 1997. Privatization and economic growth: An empirical investigation from a sample of developing market economies. *Applied Economics* 29: 161–178.

Sabirianova, K., J. Svejnar, and K. Terrell. 2005. Distance to the efficiency frontier and foreign direct investment spillovers. *Journal of the European Economic Association* 3: 576–586.

Shleifer, A., and R. Vishny. 1994. Politicians and firms. *Quarterly Journal of Economics* 46: 995–1025.

United Nations. 2004. *World investment report*. New York: United Nations.

Vickers, J., and G. Yarrow. 1985. *Privatisation and the natural monopolies*. London: Public Policy Centre.

World Bank. 1996. *World development report*. Washington, DC: World Bank.

Zinnes, C., Y. Eilat, and J. Sachs. 2001. The gains from privatization in transition economies: Is change of ownership enough? *IMF Staff Papers* 48: 146–170.

Probability

Ian Hacking

Probability denotes a family of ideas that originally centred on the notion of credibility, or reasonable belief falling short of certainty. There have arisen two quite distinct uses of this group of ideas, namely in the modelling of physical or social processes, and in drawing inferences from, or making decisions on the basis of, inconclusive data.

Modelling

We imply the most elementary of probability models when we say that a roulette is fair, meaning that the probability of the ball settling in any one segment of the wheel is equal to that of its settling in any other. Talk of fair coins or biased dice is represented in a model that is typically used to predict the relative frequency with which the possible outcomes will occur. More formal models arise from a natural abstraction and generalization of this ancient idea. In proposing a probability model for some phenomenon, one is making a claim about how some aspect of the natural, social or human world is arranged and how it behaves. Such assertions are contingent propositions that should be susceptible of empirical test. In economic theory they are typically embedded in models employing other theoretical constructs, such as utility, but the present entry is restricted to probability itself.

Inference

Probability is also used for drawing inferences from inadequate information. When combined with an assessment of utilities, it is also used for deciding what to do in the face of uncertainty. Probability is here a tool for reasoning from data and for adjusting one's beliefs or actions in the light of new evidence. Such use in reasoning is more akin to logic than to the empirical science of which probability modelling is a part.

Evidently modelling and inference are tools that need not conflict. Often they are complementary, for methods of inference or decision are often required to choose among competing models. Conversely, a probability model may, in suitable circumstances, be invaluable for drawing probable inferences and making decisions. Despite the compatibility of inference and modelling; there has been a great deal of controversy about the foundations of probability and statistics, partly but only partly arising from confusing these two distinct uses to which probability ideas can be put. The present entry will describe these foundational and conceptual issues, leaving the applications of probability to special topics treated elsewhere in these volumes.

Frequency Conceptions of Probability

It has often been urged, in order to diminish controversy, that there are two distinct and compatible conceptions of probability that perhaps deserve to be called by different names.

One idea derives from our experience of the fairly stable relative frequency of some kinds of events in repeated trials. We appear to be familiar with such phenomena, not only in manmade gambling devices, but also in nature, ranging from radioactivity to the relative frequency of births of the two sexes. A precise definition is, however, elusive.

Some writers propose that we should represent this idea in terms of an infinite sequence of trials. If the relative frequency of event *A* in a countably infinite series tends in the limit to *p*, and certain other conditions are met, then the probability of *A* is *p*. The chief additional condition is that the relative frequency of *A* should also tend to *p* in all sub-sequences that can be picked out in advance. This requirement makes a) 'gambling system' impossible (von Mises 1928); for a precise analysis of such randomness in terms of complexity see work reported in Fine (1973, ch. V).

It is often objected that limiting frequencies are too much of an idealization. Coins are seldom tossed very often, and they wear out unevenly. A chief alternative to limits is that of propensity (Popper 1959). Here a probability is taken to apply to a physical or social system, and is the tendency or disposition of the system to deliver event *A* on a single trial This may manifest itself in a stable relative frequency if sufficiently many trials are made, but the propensity is thought of as a property of the system itself. A comparison would be with the malleability of a piece of copper, taken to be a fact about the structure of the mental. It is often objected that there has never been a lucid analysis of any kind of propensity, tendency or disposition, and that talk of propensities is obscurantism. The American philosopher C.S. Peirce (1839–1914) held each idea in succession, first favouring the limiting frequency view (Peirce 1878, II, 651) and later the dispositional account (Peirce 1910, II, 664). In fact it may not be necessary to take a position on these matters. Although talk of limiting frequencies and of tendencies plays an important heuristic role in forming intuitive ideas of probability, it is not of final importance. The substantial connection between probability and frequency is provided by the limiting theorems of §8 below.

Degrees of Belief

Here the paradigm is a statement such as, 'The probability that it will be warm and sunny tomorrow is 80 %'. Of itself this cannot express a relative frequency (even if meteorological frequencies are part of the evidence for the statement), because tomorrow comes but once. The statement expresses the credibility of the thought that it will be a nice day tomorrow. There are two ways to explicate it.

First, the oldest, is the idea of rational confidence: the extent to which it is rational to be confident of hypothesis *A* (a fine day) in the light of available evidence *B*. This approach has often been called subjective, because its early proponents spoke of probability being relative in part to our ignorance and in part to our knowledge (Laplace 1795). However, it is now generally agreed that the term is misleading, for one is concerned with an objective relation between the hypothesis *A* and the evidence *B*, a probability relation analogous to the deductive relations of logic (Keynes 1921). One is concerned with reasonable degrees of belief relative to evidence, and this theory is best called a *rationalist* one.

The label 'subjective theory' should be avoided; when used, it should be applied to another account that starts with the following observation by F.P. Ramsey: a 'fundamental criticism of Mr Keynes' views,. . . is the obvious one that there really do not seem to be any such things as the probability relations that he describes' (Ramsey 1926, p. 2). This scepticism led Ramsey, B. de Finetti (1937) and L.J. Savage (1954) to develop what Savage called a theory of personal probability. Here a statement of probability is the speaker's own assessment of the extent to which he or she is confident of a proposition. It is remarkable that a seemingly subjective idea like this is arguably constrained by exactly the same mathematical rules as govern the frequency conception of probability. It is to these rules that we must now turn.

Mathematical Probability

Probability theory is a branch of the mathematical discipline known as measure theory, with the further condition that all measures are normalized, i.e. lie in the interval [0, 1]. The formal theory is open to numerous interpretations, including those mentioned in the two preceding sections.

In what follows, $P(A)$ is read as the probability of A. In the frequency interpretation, A will be an event of some kind, while in a belief interpretation, A is a proposition or hypothesis. This proposition/event nomenclature is of little moment, for we can speak of the proposition that an event occurred, or of the event that a proposition is true. In the event reading, events are represented by sets and we are concerned with an algebra of sets closed under union ($\cup$) and complementation. Intersection is denoted ($\cap$). In a propositional algebra, these correspond to disjunction, negation and conjunction. Ω denotes the sure event or certain proposition, the union of all events (or disjunction of all propositions) in the algebra.

Informally the basic laws of probability for an algebra of finitely many events (such as the six possible outcomes of a die, closed under union and complementation) are as follows:

(1) Normalization. $P(\Omega) = 1$.
(2) Non-negativity. For all events A, $P(A) = 1$.
(3) Additivity. For disjoint A, B, $P(A \cup B) = P(A) + P(B)$.

Two fundamental concepts, conditional probability and independence, may then be defined. Conditional probability is denoted by $P(A/B)$. In a frequency interpretation this is the relative frequency with which A occurs among trials on which B occurs. In a personal belief interpretation, this may be understood as the rate at which a person would make a conditional bet on A – all bets being cancelled unless condition B is satisfied. Note that in the rational belief interpretation, all probability statements are implicitly statements of conditional probability, and an axiomatization will be couched in terms of conditional probabilities.

(4) Conditional probability.

$$P(A/B) = \frac{P(A \cap B)}{P(B)} \text{for } P(B) \neq 0.$$

Finally the notion of an independent event is vital in frequency applications. Intuitively A is independent of b if the occurrence of B makes no difference to whether A occurs or not, so we expect that $P(A/B) = P(A)$. In virtue of (4), this is equivalent to:

(5) Independence. A and B are pairwise independent if and only if

$$P(A \cap B) = P(A)P(B)$$

Mutual independence of a class of n events is defined analogously.

From the very earliest days of probability calculations, speculators and gamblers were preoccupied by the fair price for a stake in a game or other transaction, such as the purchase of an annuity. Expectation, or expected value, is the formalization of this idea. Let there be a quantity X with possible values $x_1, x_2, \ldots, x_n$, and let $P(x_i)$ be the probability that X has the value x_i. In the historical origins, the x_i would be payoffs from a game and the $P(x_i)$ would be the chances of getting payoff x_i. The expectation is then defined:

(6) Expectation. $\Sigma(X) = \Sigma x_i P(x_i)$. The expectation is also called the mean value of X. The usual measure of the 'average deviation' from the mean μ is the standard deviation σ whose square σ^2 is called the variance.
(7) Variance. $\sigma^2 = E(X - \mu)^2 = \Sigma(x_i - \mu)^2 P(x_i)$.

The concepts present in (1)–(6) are clearly set forth in Huygens (1657), which takes expectation rather than probability as the primitive idea. The classic formulation of these ideas as part of measure theory is due to Kolmogorov (1933). Here a probability space is a triple $(\Omega, \mathscr{F}, P)$.

Ω: a space of mutually exclusive and jointly exhaustive events.

$\mathscr{F}$: a countable algebra of suitable subsets of Ω, that is, an algebra closed under countably infinite union and complementation.

Corresponding to (1)–(3) above we have Normalization, Non-negativity and:

(3^*) Countable additivity. For any countable sequence A_1, A_2, . . . of pairwise disjoint elements of $\mathscr{F}$,

$$P(UA_i) = \Sigma P(A_i)$$

Conditional probability, independence, expectation and variance are explained by measure theoretic generalizations of (4)–(7).

Personal Degrees of Belief and the Axioms

It is evident that both finite and limiting relative frequencies will satisfy the probability axioms. It is obscure why propensities should do so. This question is seldom addressed by those who favour that approach, but see Suppes (1973). This may be because of deeper connections between the frequency idea and the mathematical formalism; see §8 below. Here we indicate the ground for the more surprising result that, arguably, personal degrees of belief should satisfy the probability axioms.

There are two parts of the argument. (*i*) Construe degrees of belief as betting rates. (*ii*) Establish reasonable constraints on a person's set of betting rates. The best introductory exposition of these ideas is the first place they were proposed (Ramsey 1926).

Ramsey thought of a probability space as a representation of psychological states of belief. $P(A)$ stands for a person's degree of confidence in A. It is to be evaluated behaviourally by determining the least favourable rate at which this individual would take a bet on A. If the least favourable odds are, for example, 3:1, then that person's probability is $P(A) = 3/4$. Conditional probability may be explained in terms of conditional bets. Thus suppose one bets on horse A winning a race, on condition B that the horse completes the course, all bets being off if the condition B fails and the horse drops out. If a person bets 2:1 on A winning, conditional on B, then that person's conditional probability is 2/3. Ramsey was well aware that one should not expect that real betting rates should be measured to 'too many places of decimals'.

Betting is all very well, but is hardly a general psychological test, for many people go out of their way to seek or to avoid gambling, and this irrelevant factor will distort or render impossible measures of belief by betting behaviour. To get around this, imagine that a man is offered the choice of one of the two following options, at no cost to himself:

(a) He gets $1 if A occurs.

(b) He gets $1 if B occurs. If he is indifferent between the two, they are equally probable for him, while if he prefers (a) to (b), then for his personal probabilities $P(A) > P(B)$. Now if this man can generate large sets of equally probable events we can measure his probabilities to any realistic degree of precision. Suppose for example that he acts as if he thought a coin is fair, and regards any sequence of 10 outcomes of heads and tails as as probable as any other. Then he has 10! equally probable disjoint events. He can be asked about options such as:

    (a) $1 if A occurs.

    (b) $1 if heads occurs on the next two consecutive tosses. Preference for (a) indicates that his $P(A) > 3/4$; for (b), $P(A) < 3/4$, while indifference indicates that his $P(A) = 3/4$. Repeated uses of this 'risk free' technique can refine the measurement of his probabilities without any recourse to outright gambling. Suppose, then, that it makes sense to attach betting rates to a person's beliefs. Why should they satisfy the probability axioms? Why call betting rates) 'probabilities' at all? The deepest justification jointly develops probability and utility, and hence is

> outside the scope of this entry. That was the method of Ramsey (1926), which uses a perspicuous zig-zag exposition. That is, first a constraint is placed upon degrees of belief. This is used to place a constraint on utilities. This in turn is fed back into degrees of belief in the form of a further constraint. The upshot of Ramsey's paper is a full axiomatization of probability and utility. A more sophisticated version of such an approach is given in Savage (1954).

There is, however, a less compelling but easy to follow argument suggested in a throwaway phrase in Ramsey (1926), and independently developed in detail in de Finetti (1937).

De Finetti urged that a set of betting rates should be *coherent* in the following sense. It would be unreasonable or) 'incoherent' to offer betting rates on a schedule of propositions such that a clever gambler could make a profit by betting with you no matter what transpires. For a trivial example, let A = Australia retains the *America*'s cup in the next competition. Suppose a person is willing to bet on A at odds 7:3, and on the opposite, $\bar{A}$ at odds of 3:2, then the personal probabilities are $P(A) = 0.7$ and $P(\bar{A}) + 0.6$ That violates the additivity law that requires $P(A) = P(\bar{A}) = 1$.

Suppose that a gambler bets against this person on A with a stake of \$75, and against him on $\bar{A}$ with a stake of \$100. The gambler stakes \$175 and gets back \$250, regardless of who wins the next *America*'s cup. The pair of betting rates 7:3 and 3:2 are *incoherent*. A coherent set of betting rates is such that it is impossible to place a bet against them in such a way as to make a guaranteed profit in the above sense.

De Finetti proved that the necessary and sufficient conditions that a set of betting rates be coherent, is that they satisfy the probability axioms and conditional probability rule, (1)–(4) of §5 above. He does not extend this result to a fully countably additive algebra of events, because he is rigorous in his construal of personal probability. A person cannot realistically be supposed to have a structure of beliefs over such an algebra of events. However, de Finetti does extend his theory so as to allow for example for integration, in ways reminiscent of intuitionist and constructive approaches to the foundations of mathematical analysis.

De Finetti also gave a personal equivalent of independence, defined as (5) in §5 above. This was particularly important for him, because he held that in nature there are no independent trials or stable frequencies; there are only our beliefs about what will happen on individual trials. Since 'independence' is of great heuristic and mathematical power, de Finetti wished a personal surrogate for it. Corresponding to statistical independence he proposed what he called *exchangeability*. The core idea is that events that may occur in a sequence are exchangeable (in a person's belief structure) if the person is indifferent between all sequences in which the proportions of events of given kinds which occur are the same, regardless of the order in which they occur. Thus a person is indifferent between any *ABBBB, BABBB, BBABB, BBBAB* and *BBBBA;* indifferent between any of the 10 sequences containing 2 A and 3B, and so forth. Natural generalizations of this idea lead to a powerful theory of exchangeability (Diaconis and Freedman 1980).

Bayesianism

The probability axioms have an immediate consequence much used in the theory of personal probability. Let $A_1, \ldots, A_n$ be a set of mutually exclusive and jointly exhaustive events. By (4) of §5, for each $i \leq n$,

$$P(B)\,P(A_i/B) = P(A_i)P(B/A_i).$$

Since

$$B = \cup\left(A_j \cap B\right)$$
$$P(B) = \sum P\left(A_j \cap B\right) = \sum P\left(A_j\right)P(B/A_i).$$

The first and third lines imply that

$$P(A_i/B) = \frac{P(A_i)P(B/A_i)}{\sum P(A_j)P(B/A_j)}. \qquad (8)$$

(8) is often called Bayes' theorem, due to its tenuous connection with Bayes (1763). As it is a trivial deduction it is better called Bayes' rule. It is of course valid in any interpretation, but it is of serious interest only from a belief point of view. For suppose that the partition $A_1, \ldots, A_n$ is an exhaustive set of mutually exclusive hypotheses of interest, and that B is evidence bearing on the hypotheses. Suppose further that a person has, in the light of prior knowledge, a distribution of belief over the A_i, represented by $P(A_i)$ for each i. Call this the *prior* probability distribution. Let the A's be of such a sort that for each A_i, $P(B/A_i)$ is defined. This is called the *likelihood* of getting B, if A_i is true. For example, let A_1 state that the outcomes of a die are equiprobable, and A_2, that there is bias with $P(6) = 0.3$. Then the likelihood of 6 on A_1 is 1/6, and A_2 is 0.3.

Now we may ask, in order to be coherent, how should a person incorporate a new piece of evidence B into a prior probability distribution over the hypotheses A_i? A plausible answer following Bayes' rule is that the 'posterior distribution' (in the light of B) should be the same as the (prior) conditional probability distribution $P(A_i/B)$. Schematically,

Posterior probability ∝ prior probability × likelihood

The constant of proportionality is as in (8) above.

It is argued that this provides a model of reasonable learning for experience. A person at some stage has a purely personal prior probability distribution over some range of possibilities. This is subject only to the constraint of coherence. However, it is urged, coherence entails a uniquely reasonable way to adjust one's probabilities in the light of new information. On learning new B, one should move from prior probabilities $P(A_i)$ to posterior probabilities equivalent to $P(A_i/B)$. If we spoke of personal probabilities at a time, indicated by a superscript t, and if between t and t' we learn just: B, then

$$P^{t'}(A_j) \text{ should be } P^t(A_i/B). \qquad (9)$$

This is an additional postulate that does not follow from the probability axioms, and which is peculiar to the personal interpretation of probability. Nothing in the coherence argument entails that probabilities should be adjusted across time in accord with (9). Attempts to justify statements such as (9) have been interesting but inconclusive (Diaconis and Zabell 1982). Assertion (9) should be regarded as a pragmatic postulate for the use of personal probabilities. No one has proposed a significantly different and seriously better general rule for personal learning from experience.

Every interpretation of probability appears to require a pragmatic postulate, although what is required is different in each case. Consider the theory of rational belief, in which the conditional probability $P(A/B)$ is interpreted as a logical relation between B and A. This is supposed to be a unique constant determined by something analogous to deductive logic. Here there is no need for a time-spanning postulate such as (9), for in a theory of rational belief, all probabilities are conditional and are held to be uniquely defined. How can such purely logical relations serve as 'the very guide of life' (Butler 1736, Introduction) that one expects from an applied theory of probability? How can a collection of logical relations help with predictions and decisions? One cannot reply that if the total available evidence bearing on A is B, then the rational probability of A is $P(A/B)$, because on the theory of rational probability, 'the probability of A' makes no sense, all probabilities being relative.

One requires a pragmatic postulate similar to what has been called *the requirement of total evidence* (Carnap 1950, p. 211). Let $A_1, \ldots, A_n$ be a partition of states of affairs, and B a further proposition. Let $P(A_i/B)$ be a rational probability function defined for fixed B. Let U_i be the value or utility if A_i obtains. Then the rational expectation on B is, following (6) of §5,

$$\sum U_i P(A_i/B).$$

A pragmatic posultate would state: if B is the total available evidence relevant to members of the partition Ai, then act so as to maximize the rational expectation on B. Since it is doubtful whether there exist rational probability functions, and since the notion of total available evidence is utterly obscure, such a postulate is of purely academic interest, intended to illustrate a feature of some traditional reflections on probability.

The term Bayesian is at present commonly used for a personal theory of probability that makes heavy use of Bayes' rule and (implicitly) a pragmatic postulate. But is should also include rationalist approaches such as those of Jeffreys (1939) which also make extensive use of Bayes' rule.

Limit Theorems

Bayes' rule is a theorem valid under any interpretation of probability but whose interest is largely confined to degree of belief approaches. We now turn to a fundamental body of work whose immediate application is more evident for frequency approaches. Arguably this work, which results in a series of limit theorems, establishes the essential connection between probability and finite relative frequencies. It begins with a result proved by Jacob Bernoulli around 1695 (Bernoulli 1713).

Bernoulli intended his investigations to contribute to an understanding of statistical inference, but they are also important for the very conception of probability.

Regardless of the glosses of §3 above, in terms of limiting frequency or propensity, the intuition underlying an abstract frequency conception of probability is this: if A occurs k times in n independent and 'identical' trials, then, if n is large, k/n should be close to the probability p to A. This is the content of Bernoulli's theorem. In its weak form it states that for any small 'error' ϵ,

$$P\left(\left|\frac{k}{n}-p\right|<\varepsilon\right)\rightarrow 1 \text{ as } n\rightarrow\infty. \qquad (10)$$

A stronger form asserts that for any error ϵ and small probability δ there is a number N such that for $n > N$,

$$P\left(\left|\frac{k}{n}-p\right|<\varepsilon\right)\rightarrow 1-\delta. \qquad (11)$$

This applies only to 'identical' trials on which the probability is $P(A) = p$ for each trial. S.-D. Poisson coined the term 'law of large numbers' for a generalization (Poisson 1837). Suppose that the probability of A on successive trials is not constant, but only that there is a suitable regular probability distribution for values of $P(A)$. In Poisson's examples, one would have a sequence of urns, each with a different proportion of black as opposed to white balls. $P(A)$ would be the probability of drawing a black ball from a given urn, and there would be a probability distribution over the proportions of balls in successive urns. Poisson established that a result analogous to (11) holds, in which p is replaced by the mean probability of $P(A)$. The term 'law of large numbers' is now widely used to apply to all results of this type, including Bernoulli's original (10).

A continuation of this result began with work by A. de Moivre in 1732 and fully developed by P.S. Laplace and C.F. Gauss around 1800 (Stigler 1986). It is the beginning of the well known Gaussian or Normal probability distribution with its familiar 'bell shaped curve' once known simply as *the* curve of probability.

Bernoulli had been able to place crude upper bounds on the probability that on n trials the proportion of A occurring should be within ϵ of p. De Moivre addressed the more general question of the form of the histogram of k/n (for $k = 0, \ldots, k = n$) as n grows without bound. Let $\Phi(x)$ be a cumulative probability distribution for a real-valued variable x: thus $\Phi(x)$ is the probability of the variable taking a value less than or equal to x. The Normal distribution with mean zero and variance 1 is

$$\Phi(x)=\frac{1}{\sqrt{(2\pi)}}=\int_{-\infty}^{x}\exp\left(-\frac{1}{2}y^2\right)\mathrm{d}y.$$

In repeated independent trials with fixed probability $P(A) = p$, the following holds. For any fixed $a < b$, as $n \to \infty$,

$$P\left\{a \leq \left|\frac{k}{n} - p\right| \left[\frac{n}{p(1-p)}\right]^{1/2} \leq b\right\} \\ \to \Phi(b) - \Phi(a). \qquad (12)$$

This establishes with greater precision the connection between probability and stable long-run frequency. It also shows how the probability distribution associated with the simplest chance device, namely coin tossing, relates to the Normal distribution which, for mean μ and variance σ^2 is:

$$\frac{1}{\sqrt{(2\pi)}\sigma}\int_{-\infty}^{x} \exp\left[-(y-\mu)^2/2\sigma_2\right] \mathrm{d}y.$$

Proofs of (12) and increasingly general theorems of that type were developed by the Petersburg school of mathematicians in the latter part of the nineteenth century: P.L. Chebyshev, A.A. Markov and A.M. Lyapunov (see Maistrov 1967). Lyapunov gave one form of the *central limit theorem*. Consider a sequence of mutually independent variables $x_1, \ldots, x_r$ with a common distribution. Suppose that the mean and variance of x_i exist, and are μ and σ^2. Let $k = x_1 + \ldots + x_n$. Then for every fixed ϵ as $n \to \infty$,

$$P\left\{\left|\frac{k}{n} - \mu\right| \left(\frac{n}{\sigma}\right)^{1/2} < \varepsilon\right\} \to \Phi(\varepsilon). \qquad (13)$$

From a conceptual point of view, central limit theorems should be regarded as a culmination of a series of results that explicate the fundamental frequency conception of probability. They also illustrate the power of the probability axioms.

Application of Such Probabilities

Although the limit theorems lay bare the intuitive connection between probability and frequency, there remains a question of application to a single and unique event, such as the next outcome of a roll of a die. If we think of probability in terms of relative frequencies, how can it bear on a particular toss? We may believe that the probability of 6 with a die is 1/6, and believe, thanks to the limit theorems, that in a long run the relative frequency of sixes will usually be close to 1/6. But if we are to judge only the next outcome, for example for purposes of making a bet, why should the fraction 1/6 be of special interest? Once again it appears that, just as for a personal probability interpretation, we require a pragmatic postulate, this time in order to decide what to do *next*, in other words, once again in order to make probability 'a very guide of life'. This has in effect been proposed by many writers, for example by Reichenbach (1949, §72), who speaks of single case 'posits' in connection with individual cases.

There is, however, an additional problem, often called the problem of the reference class. A particular future event may be a member of several classes, each of which is associated with a stable relative frequency. A man may be both a heavy smoker and a jogger. Probabilities of living to the age of 60 may be known for smokers of his age, and for joggers of his age, but neither in itself tells us what to expect of this individual. It is commonly proposed that an individual case should be referred to the smallest discernible class for which a frequency is known. Then a pragmatic postulate would instruct one to act so as to maximize expectations relative to smallest discernible reference classes.

This postulate is curiously similar, in certain respects, to the pragmatic postulate needed for Bayesian learning from experience. Even if expectation may seem a good guide in life in situations where one is to make a series of successive decisions or take a series of successive gambles, there is no compelling reason for employing expectations in a particular case. All the same (just as with (9), the temporal use of Bayes' rule) no alternative yet proposed has, in general, any desirable features at all.

Non-Quantitative Probability

The classical approach to probability from Huygens (1657) to the present has measured probabilities by rational fractions or real numbers.

A minority of workers has rejected this as unrealistic. Thus Keynes supposed that comparisons of probability are often feasible when quantitative measures are not (Keynes 1921). He had the rationality approach, in which all probabilities are conditional, and he held that the fundamental form of probability statements is:

$$P(A/B) \leq P(C/D).$$

This gives rise to a lattice of partially ordered probabilities that has proved attractive from various points of view. Some personalists hold that we can usually only compare our personal degrees of confidence, not measure them. A few frequentists have held that stabilities in nature are seldom secure enough to guarantee quantitative long-run frequencies, but that we have many comparisons furnished by nature. For a survey of these approaches, see Fine (1973, ch. II).

Inference and Modelling

We now return to the chief uses of probability stated in §§1, 2, namely inference and modelling. It might seem as if inference would naturally employ a belief-oriented conception of probability, and that modelling would be via frequency or propensity. This would imply a quite eclectic approach to interpretations of probability, as is common among day-to-day consumers of probability mathematics. However, a majority of workers on foundations have been dogmatic, arguing that we should use one and only one interpretation of probability for all purposes of interest.

Thus a majority of authors who present theories of statistical inference favour a frequency interpretation. They hold that the only legitimate tool for objective public discussion must be a frequency oriented idea. An account of personal probability is too subjective for scientific inference or public decision making.

Conversely, many adherents to a degree of belief approach hold that there just do not exist any objective propensities or frequencies in macroscopic nature. (Some but not all admit a place for them in microphysics.) Such writers try to re-express everything valuable in a frequency approach in terms of personal belief. De Finetti's exchangeability of §4 above was intended to provide a subjective surrogate for the notion of objective independence (which de Finetti thinks does not exist in nature).

Such dogmatic personalism precludes any use of modelling of natural processes by frequency-like structures, and reduces all reasoning about nature, where we have incomplete information, to operations with coherent personal probabilities. Inference, on the other hand, is no problem for the personalist, who augments the probability axioms with a (usually implicit) pragmatic postulate (9) as in §5 above. All inference is by Bayes' rule or by a more sophisticated version of that.

Conversely, the dogmatic frequentist has no problem with modelling natural processes by probability structures interpreted in terms of stable frequencies or propensities. But whereas in a sense the personalist has no theory of statistical inference (all inference being by Bayes' rule), the frequentist has not one but several competing theories of considerable conceptual difficulty. These will now be briefly described.

Inferring by Frequency

There has evolved an enormous battery of techniques for testing statistical hypotheses, estimating statistical parameters, designing experiments and making decisions. There is less agreement on the foundations for these techniques. From 1920 until his death in 1962, R.A. Fisher was a prolific source of such fundamental ideas as significance tests, maximum likelihood estimators, randomized experimental design, sufficient statistics, information, and a host of others. In part because he favoured many different and non-equivalent uses of probability, it is not possible to give a brief simple account of what he took to be the basics. The opposite is the case for a later and influential pair of workers, Jerzy Neyman and E.S. Pearson, whose theory will be sketched first.

Neyman was a dogmatic frequentist who held that it is never possible to make a probability statement about a particular event or hypothesis.

Inductive *inference* is, he said, impossible, and is to be replaced by a theory of inductive *behaviour*. We can at best choose a policy that has desirable ‘operating characteristics’. In the case of testing one statistical hypothesis H against a family of others, we require a method that seldom rejects H if it is true, (say 1 per cent of the time), and, given that constraint, usually rejects H if it is false. In most practical situations this goal cannot be uniformly achieved, but Neyman and Pearson introduce other operating constraints so as to design unique tests. Likewise in estimating that an unknown parameter lies in a certain interval on the real line, one requires that on repeated experiments the method of estimation would yield an interval that includes the true parameter most of the time. A 99 per cent confidence interval is derived by a method that is correct 99 per cent of the time (and is subject to other constraints to ensure uniqueness). In a particular case one *cannot* assert that there is 0.99 probability that the unknown parameter is within the bounds of the estimate. One can say only that the interval was derived by a method that is usually right.

Fisher regarded statistical inference as primarily a procedure for data analysis, for maximizing information obtained from experiment, and for producing intelligible informationpreserving summations of data that would otherwise be too complex to understand. He thought of the various significance levels (analogous to confidence levels) as convenient standards by means of which experimental workers could judge each others’ results. Thus in considering a treatment (of a field with fertilizer, of patients with medicine, of an economy with a rise in interest rates) one wants to know if the treatment is efficacious. Thus one tests the ‘null hypothesis’ by making a probability model of the hypothesis of no effect. The result of an experiment is significant at the 1 per cent level if, on the model of the null hypothesis, an effect at least as large as that observed would occur with probability less than or equal to 0.01. Otherwise the result is judged not to be significant at the 1 per cent level. In the event that the treatment is judged significant, one is not obliged to reject the null hypothesis. The result of a significance test is always of the following logical form: either something very unusual has occurred by chance, as will from time to time happen, or else the null hypothesis is false and the treatment is efficacious (Fisher 1956, ch. III).

Such a piecemeal approach, in which a statistical report summarizes a situation and leaves other experimenters to make up their own minds, is in apparent contrast with the regimented policies derived by the Neyman–Pearson theory. For Fisher’s chief papers, see Fisher (1950); for the joint work of Neyman and Pearson, see Neyman and Pearson (1967).

Most ordinary practitioners do not draw firm lines between the two approaches. There remain significant practical differences. For example, in certain Fisherian analyses based on likelihood (see definition in §5 above) it makes no difference whether the length of an experiment has been determined in advance, or whether the experimenter decides in the course of the experiment when to quit. On the Neyman theory, such optional stopping completely changes the analysis of the data. Likewise there are striking contrasts between frequency theories on the one hand – be they those of Fisher or Neyman – and Bayesian belief theories on the other. The former not only regularly incorporate randomization into the design of experiments, but hold it to be an essential procedure for increasing the amount of information derived from an experiment. On a Bayesian account, however, randomization is of no value and may actually lead to a loss of information. It remains the case that almost all experimenters favour randomization, and in the case of human subjects, urge double blind experimentation when practicable.

Probability and Economics

Expectation, defined in terms of utility and probability, is an economic concept. The first book on the probability calculus, one which set the pace in the early days, defined probability in terms of expectation (Huygens, 1657), and this practice continued until at least the time of Bayes (1763). Hence one would anticipate a longstanding relationship between probability and economics. This has not been the case. French pre-revolutionary physiocrats thought that probability should play a central role in ‘moral science’, which would include what we now call economics. However, this had little direct influence, even if through the mediation of the Marquis de Condorcet it aroused the early interests of that greatest of probability mathematicians, Laplace.

Only during the 19th century did probability become a working tool of a sizable number of sciences, disciplines and practices. The complete assimilation of probability to almost every topic has occurred only very recently. (For cross-disciplinary studies, suggesting the differential adoption of probability tools and techniques, see Daston et al. 1987.) Despite the fact that maximization of utility has long been an economic adage or even tautology, economics has been one of the slower disciplines to be penetrated by probabilistic thinking.

Economists have usually been eclectic in their use of probability. F.Y. Edgeworth, author of the article *Probability* in the original *Palgrave*, noted in his longer essay for the *Encyclopaedia Britannica* that a belief approach to probability might be rejected on the ground that it was ‘merely psychological’. He goes on to mention a limiting frequency approach. But, he continues, ‘these views are not so diametrically opposed as might at first appear’ (Edgeworth 1911, p. 377). Many economists would echo his words today.

Some, however, have strongly favoured only one approach to probability. The most notable example is Keynes. Although he became less dogmatic about probability later in life, at the time of his major contribution to the field (Keynes 1921) he provided the classic statement of the rationalist approach to probability, and also was far ahead of his time in urging that comparative probabilities are fundamental. F.P. Ramsey, who in his very short life had commenced important contributions to economic theory, is of course the founder of the modern personalist approach, and he was the first to see how probability and utility can be jointly axiomatized as concepts that are integrally and necessarily connected.

The theory of games and economic behaviour, set forth in its modern form by von Neumann in a work with that title, settles on the ‘perfectly well founded interpretation of probability as frequency in long runs’ (Von Neumann and Morgenstern 1944, p. 19). In a footnote it is said that one may instead axiomatize probability and preference jointly. This was done explicitly by L.J. Savage (1954), who had been one of von Neumann’s wartime assistants, and who has provided us with the standard exposition of personal probability, one which has, as one of its consequences, von Neumann’s theory of utility. The present entry has emphasized controversies about the interpretation and application of probability ideas; it must conclude by stating that despite differences in foundation, a great many divergences are washed away in the routine of day-to-day application and derivation.

See Also

- ▶ Induction
- ▶ Likelihood
- ▶ Ramsey, Frank Plumpton (1903–1930)
- ▶ Random Variables
- ▶ Statistical Inference
- ▶ Subjective Probability

Bibliography

Bayes, T. 1763. An essay towards solving a problem in the doctrine of chances. In *Studies in the history of statistics and probability*, ed. E.S. Pearson and M.G. Kendall. London: Griffin, 1970.

Bernoulli, J. 1713. *Ars conjectandi*. Basle.

Butler, J. 1736. *The analogy of religion*. London.

Carnap, R. 1950. *Logical foundations of probability.* Chicago: University of Chicago Press.

Daston, R., M. Heidlberger, and L. Krüger (eds.). 1987. *The probabilistic revolution*, 2 vols. Cambridge, MA: Bradford Books.

de Finetti, B. 1937. *Foresight: Its logical laws, its subjective sources*. Trans. from the French in *Studies in subjective probability*, ed. H.E. Kyburg Jr., and H.E. Smokler. New York: Wiley. 1964.

Diaconis, P., and D. Freedman. 1980a. De Finetti’s generalizations of exchangeability. In *Studies in induction logic and probability*, vol. II, ed. R.C. Jeffrey. Berkeley/ Los Angeles: University of California Press.

Diaconis, P., and D. Freedman. 1980b. De Finetti’s theorem for Markov chains. *Annals of Probability* 8(1): 115–130.

Diaconis, P., and S. Zabell. 1982. Updating subjective probability. *Journal of the American Statistical Association* 77: 822–830.
Edgeworth, F.Y. 1911. Probability. In *The Encyclopaedia Britannica*, vol. XXII, 11th ed. New York: Encyclopaedia Britannica.
Fine, T. 1973. *Theories of probability*. New York: Academic Press.
Fisher, R.A. 1950. *Contributions to mathematical statistics*. New York: Wiley.
Fisher, R.A. 1956. *Statistical methods and scientific inference*. Edinburgh: Oliver & Boyd.
Huygens, Chr. 1657. *Ratiociniis in aleae ludo* (Reasoning in a game of chance). In *Oeuvres complètes*, ed. C. Huygens. The Hague: M. Nijhoff, 1888–1950, includes the original Latin, Huygens' original Dutch, and French translation in Vol. 14.
Jeffreys, H. 1939. *Theory of probability*, 3rd ed, 1961. Oxford: Clarendon Press.
Keynes, J.M. 1921. *A treatise on probability*. London: Macmillan.
Kolmogorov, A.N. 1933. *Foundations of the theory of probability*. Trans. from the German, New York: Chelsea, 1950.
de Laplace, P.S. 1795. *A philosophical essay on probabilities*. Trans. from the French. New York: Dover. 1951.
Maistrov, L.E. 1967. *Probability theory: A historical sketch*. Trans. from the Russian, New York: Academic Press, 1974.
Neyman, J., and E.S. Pearson. 1967. *Joint statistical papers*. Cambridge: Cambridge University Press.
Peirce, C.S. 1878, 1910. *Collected papers of Charles Sanders Peirce*, ed. C. Hartshorne., and P. Weiss. Cambridge, MA: Harvard University Press, 1965
Poisson, S.-D. 1837. *Recherches sur la probabilité des jugements en matière criminelle et en matière civile*. Paris.
Popper, K.R. 1959. The propensity interpretation of probability. *British Journal for the Philosophy of Science* 10: 25–42.
Ramsey, F.P. 1926. Truth and probability. In *Foundations: Essays by F.P. Ramsey*, ed. D.H. Mellor. London: Routledge & Kegan Paul, 1978.
Reichenbach, H. 1949. *The theory of probability*. Berkeley/Los Angeles: University of California Press.
Savage, L.J. 1954. *The foundations of statistics*. New York: Wiley.
Stigler, S. 1986. *The history of statistics: The measurement of uncertainty before 1900*. Chicago: University of Chicago Press.
Suppes, P. 1973. New foundations of objective probability: Axioms for propensities. In *Logic, methodology and philosophy of science*, vol. IV, ed. P. Suppes et al. Amsterdam: North-Holland.
Von Mises, R. 1928. *Probability, statistics and truth*. Trans. from the German, London: Allen & Unwin, 1957.
Von Neumann, J., and O. Morgenstern. 1944. *Theory of games and economic behavior*, 3rd ed. Princeton: Princeton University Press, 1953.

Procurement

Yeon-Koo Che

Abstract

Firms and government agencies rely increasingly on goods and services procured from outside suppliers. How to assure desired quality at a minimal cost in the procurement is often challenging and warrants carefully devised contracting policies. This article reviews several problems arising in procurement and policies designed to remedy them.

Keywords

Adverse selection; Auction contests; Auctions; Collusion; Corruption; Cost-plus contracts; Fixed-price contracts; Mechanism design; Option contracts; Outsourcing; Procurement; Quality assurance; Reputation; Risk aversion; Risk sharing; Scoring auctions; Virtual cost

JEL Classification

D6

Virtually all businesses, both public and private, rely on procurement of numerous goods and services, ranging from such routine jobs as food and custodial services to the complex job of building high-tech fighter jets and high-speed train systems. Rapid progress of communication technologies – most notably the emergence of the Internet – has made outsourcing both cheaper and more efficient, thus altering the traditional boundaries of 'make-or-buy' decisions by many firms and government agencies in favour of more outsourcing. Thus, designing efficient mechanisms for procuring goods and services has become ever more important.

Procurement of standardized parts and services is relatively straightforward as a competitive market or standard bidding would produce an efficient outcome. In many procurement settings, however, the quality of the procured job is an important concern, and it is not easy to assure the desired level of quality, since a high-quality job may entail a cost that is unknown or privately known by the supplier or require his special effort that is not observable to the buyer, or the quality of the job provided is not easy to verify or is simply unobservable to the buyer. The present article reviews some of the answers economics research has provided for optimal responses to these procurement problems.

Contractible Quality

When quality is verifiable, the terms of the contract can be made contingent on the quality of the system. If the cost associated with delivering the job is unobservable, it presents two problems. First, ensuring the supplier's participation may require paying more than the true cost, that is, information rents to elicit the cost, so the quality level must be decided based on the overall cost paid to the supplier. Second, the buyer must identify the supplier who can deliver the good at the minimal cost to her. We sketch the method for finding the optimal mechanism that deals with these issues.

To begin, suppose a buyer derives utility of $v(q) - t$ when she procures a job of quality $q \in \mathbb{R}_+$ and pays $t \in \mathbb{R}_+$, where the gross surplus function, $v : \mathbb{R}_+ \mapsto \mathbb{R}_+$, is strictly increasing, differentiable and strictly concave. Suppose there are $n \geq 1$ potential suppliers. Suppl ier $i \in N := \{1,\ldots,n\}$ can deliver quality at unit cost of θ_i, which is drawn from $\left[\underline{\theta}, \overline{\theta}\right] =: \Theta$ according to the cumulative distribution function $F_i(\cdot)$ which has a positive density over $\left(\underline{\theta}, \overline{\theta}\right)$. Assume also that $\theta + \frac{F_i(\theta)}{f_i(\theta)}$ is non-decreasing in θ. A supplier i receives t - $\theta_i q$ from a contract that pays him t for delivering q.

If the suppliers' costs are observable, then the procurer's decision will be straightforward. She can pick the most efficient one, $i = \arg\min_{i \in N} \{\theta_i\}$, and have him deliver the job at cost, so she will pick the first-best quality, $q_i^{**}(\theta_i) \in \arg\max_{q \in \mathbb{R}_+} v(q) - \theta_i q$.

When the suppliers' costs are unobservable, it is not possible to procure at the actual cost, since suppliers can pretend to have higher than actual cost. Nor is it easy or necessarily desirable to pick the least-cost supplier, as will be seen. To illustrate the optimal procurement decision, suppose first there is only one potential supplier, $n = 1$. By the revelation principle, there is no loss in restricting attention to a direct revelation contract that determines the quality and the payment, $\{(q_i(\theta), t_i(\theta))\}_{\theta \in \Theta}$, as a function of the cost reported by the supplier.

The optimal contract $\left(q_i^*(\cdot), t_i^*(\cdot)\right)$ must solve

$$\max_{(q_i(\cdot),\, t_i(\cdot))} \int_{\underline{\theta}}^{\overline{\theta}} [v(q_i(\theta)) - t_i(\theta)] dF_i(\theta) \qquad ()$$

subject to

$$U_i(\theta) := t_i(\theta) - q_i(\theta) \geq 0, \forall \theta \in \Theta \qquad \text{(IR)}$$

$$U_i(\theta) \geq t_i\left(\widetilde{\theta}\right) - \theta q_i\left(\widetilde{\theta}\right), \forall \theta, \widetilde{\theta} \in \Theta, \qquad \text{(IC)}$$

where (IR) and (IC) ensure, respectively, the supplier's participation and his incentive to report truthfully his type.

By the well-known method, (IR) and (IC) constraints can be simplified to a pair of conditions:

$$q(\cdot)\,\text{is}\,\text{non}-\text{increasing}. \qquad \text{(M)}$$

and

$$U_i(\theta) = \int_{\theta}^{\overline{\theta}} q\left(\widetilde{\theta}\right) d\widetilde{\theta}, \qquad \text{(Env)}$$

or equivalently

$$t_i(\theta) = \theta q_i(\theta) + \int_{\theta}^{\overline{\theta}} q\left(\widetilde{\theta}\right) d\widetilde{\theta}. \qquad \text{(Env}'\text{)}$$

The constraint (M) will be seen not to bind, so it can be ignored. Substituting (Env′) into the objective function of [P] and switching the order of expectations yield

$$\int_{\underline{\theta}}^{\overline{\theta}} [v(q_i(\theta)) - J_i(\theta) q_i(\theta)] dF_i(\theta),$$

where $J_i(\theta) := \theta + \frac{F_i(\theta)}{f_i(\theta)}$ is the so-called 'virtual' cost. The additional cost $\frac{F_i(\theta)}{f_i(\theta)}$ reflects the additional rents that must be given away to the types more efficient than θ when its quality is raised marginally. To see this more intuitively, suppose the quality for type θ is raised by Δq towards the efficient level. Then, there is an efficiency gain (to be captured by the procurer) of $[v'(q) - \theta]\Delta q f(\theta)$. At the same time, the raising of quality enables each type $\theta' < \theta$ to command extra rents of Δq by mimicking (or choosing the contract intended for) type θ, so the same amount must be given to them to dissuade them from doing so. Since the measure of those types is $F(\theta)$, the marginal cost of quality increase is $F(\theta)\Delta q$. The optimal quality $q_i^*(\theta)$ balances these two marginal effects, so $v'(q) - \theta \frac{F_i(\theta)}{f_i(\theta)}$ if $q_i^*(\theta) > 0$, or more generally

$$q_i^*(\theta) \in \arg \max_{q \in \mathbb{R}_+} [v(q) - J_i(\theta)q].$$

Clearly, $q_i^*(\theta) < q_i^{**}(\theta)$, for $\theta > \underline{\theta}$, whenever $q_i^{**}(\theta) > 0$. In other words, it is optimal for the buyer to choose less than the first-best quality. In particular, the buyer may not procure at all even though procuring is socially efficient, for instance when $\theta < v'(0) < J_i(\theta)$. In practice, the optimal procurement policy can be implemented by a *menu* of quality-transfer pairs, $\left(q_i^*(\theta), t_i(\theta)\right)$, or by a *nonlinear pricing scheme* $\tau(q) := t_i\left(q_i^* - 1(q)\right)$.

Now suppose $n \geq 2$ so there are multiple candidate suppliers. The selection of the supplier, which can be studied using the same mechanism design approach (see Myerson 1981; Laffont and Tirole 1987; McAfee and McMillan 1987; Riordan and Sappington 1987), extends the above insight naturally. What ultimately matter to the buyer are suppliers' virtual costs, not their actual costs. Hence, the supplier with the lowest virtual cost, $i \in \arg\min_{j \in N}\{J_j(\theta_j)\}$, must be selected, and the selected supplier must choose the 'downward distorted' quality level, $q_i^*(\theta)$. If supplier i has *ex ante* higher cost than j, say in terms of conditional stochastic dominance: $\frac{F_i}{f_i} < \frac{F_j}{f_j}$, then the optimal selection rule favours i. Favouring the 'underdog' can be seen as a way of *handicapping* the top dog to make him compete more aggressively.

When the suppliers are *ex ante* symmetric, that is, $F_i = F_j$ for $i \neq j$, then the optimal selection is also efficient, and the optimal procurement policy can be implemented by the so-called *scoring auction* (see Che 1993). Specifically, there is a quasi-linear scoring function,

$$S(q, t) := v(q) - \varDelta(q) - t,$$

for some $\varDelta(\cdot)$ increasing, that implements the optimal outcome if the suppliers are asked to make two-dimensional bids, (q, t), and the supplier who achieves the highest score according to $S(q, t)$ is selected to produce his proposed quality and receive his payment. The term $\Delta(q)$ serves as a penalty against 'quality bid' so as to implement the downward distortion feature of the optimal contract. The scoring auction resembles the procedures used in the procurement of weapons, transportation, construction, and a multitude of other goods and services. *Quasi-linear scoring auctions* are analytically tractable and can implement a broad range of outcomes, even when the quality is multidimensional (so q is a vector of attributes) and the suppliers may have heterogeneous costs with these attributes (Asker and Cantillon 2004). (A quasi-linear scoring auction may not implement the optimal direct revelation mechanism for the buyer if the suppliers have multidimensional costs, but it does implement the socially efficient quality mix. See Asker and Cantillon 2005.)

In many procurement settings, the monetary expenditures are observable, but the suppliers' inherent capabilities as well as their effort to reduce the cost may not be observable. In this case, how a supplier's cost should be reimbursed becomes an important issue. A *fixed-price contract* that pays the same price to the supplier regardless of his realized cost provides a strong incentive for cost-reduction but requires the supplier to bear the risk of cost shocks. By contrast, a *cost-plus contract*, which reimburses the supplier's cost fully, provides weak incentives for cost-reduction effort but imposes no risk on the part of the supplier. McAfee and McMillan (1986) show a mixture of the two contract forms — that is, a *partial reimbursement rule* — to optimally balance the trade-offs between the cost reduction incentive, adverse selection and risk sharing. (Laffont and Tirole 1986, obtained a similar result without risk aversion of the agent.) Bajari and Tadelis (2001) focus on the trade-off between the cost reduction incentive and *ex post* renegotiation inefficiencies, and study how the complexity of the procurement job affects the choice of contract form. They argue that fixed-price contracts are optimal for standard jobs, whereas cost-plus contracts are optimal for complex projects. Empirical findings on the contract choice appear consistent with this latter finding (Crocker and Reynolds 1993; Corts and Singh 2004; Bajari et al. 2002).

Uncontractable Quality

Often the quality enjoyed by the buyer is unobservable to the supplier and/or unverifiable to the court, so it is difficult to contract on it *ex ante*. Book publishing, advertising, film production, development of new (such as pharmaceutical) technologies, procurement of new weapons systems and hiring new talents all involve some difficulty in specifying the quality of jobs. While *ex post* signals about quality are often available after the procurement (for example, sale of a book or of an advertised product), the fixed cost associated with procurement may be so high that quality assurance is needed before full-scale production begins. We discuss several methods for assuring quality.

To illustrate, suppose a buyer values the good at q if a supplier makes an effort $c(q)=\frac{1}{2}q^2$. It would be ideal for the buyer to obtain the quality $q^*=1$ at the price of $c(q^*)=\frac{1}{2}$. If quality is unverifiable, it would be difficult to specify contractually the level of quality. The buyer would argue that the quality provided is lower than specified, and the supplier would argue the opposite. In any case, the supplier would have little incentive to provide high quality, since there would be little reward for it.

A simple *option contract* can solve the problem of unverifiable quality. Suppose the buyer signs a contract that requires the supplier to pay a (non-refundable) upfront fee of q^* - $c(q^*)$=$\frac{1}{2}$ to the buyer and gives the buyer an option either to accept the good at the price of $p=q^*=1$ or to reject it at no penalty. If the supplier produces quality of q, then the buyer would receive $q-\frac{1}{2}$ from accepting the good and $\frac{1}{2}$ from rejecting the good. Hence, the buyer will purchase the good if and only if $q-\frac{1}{2}\geq\frac{1}{2}$, or the quality is at least $q^*=1$. Knowing this, the supplier will produce $q^*=1$. The supplier has the incentive to provide adequate quality since the buyer has an option to reject the good if the quality is not to his liking. These option contracts, known by such names as *purchase upon approval* and *delivery-contingent* contracts, are common in situations where quality assurance is important (Taylor 1993; Che and Hausch 1999). For instance, advertising agencies must often develop acceptable pilot campaigns before they are paid in full; real estate agencies and other brokers are typically not paid until they find an acceptable match between buyers and sellers; book publishers often reserve the right to recover an advance in the event that they find the book unacceptable. The *up-or-out* contracts well-known for academic tenure and law partnerships are a form of an option contract, presumably motivated to deal with the 'unverifiable quality' problem (Kahn and Huberman 1988).

A problem with the option contract is that it requires the supplier to pay an upfront fee, that is, to buy in. Often, suppliers have limited liability or are liquidity constrained, which can make the option contract infeasible. In the above example, for instance, if the supplier cannot be induced to 'buy in', the buyer will receive zero net surplus. (This is attributable to the deterministic nature of quality. If quality is stochastic, the quality accepted will generally exceed the option price. Even with stochastic quality, however, the option contract will be of limited value to the buyer if the suppliers cannot buy in.) This problem can be solved by a *pilot/research contest*, that is, by having multiple suppliers compete for a reward. To be concrete, suppose the buyer invites two suppliers, each with the same technology described 25 above, and suppose the buyer promises a fixed prize $P=\frac{8}{25}$ (which turns out to be the optimal level) for the supplier who offers the higher quality. It is then equilibrium behaviour for each supplier to randomize quality over $[0, 4/5]$ according to the CDF, $F(q) = \frac{25}{16}q^2$, yielding a surplus of $\frac{8}{25}$ to the buyer. (If the other supplier follows the randomization strategy, a given supplier receives a payoff of $F(q)P - c(q) = \frac{1}{2}q^2 - \frac{1}{2}q^2 = 0$, when choosing $q \in [0, 4/5]$, so randomizing according to F is a best response.)

Fixed-prize contests have been used for developing new innovations, some historically important, such as the *longitude technology* and the *steam engine design*. But recent procurement contests have allowed suppliers some freedom in specifying their own rewards. For instance, most of the defence procurement competitions as well as grant competitions allow suppliers to adjust the size of their prizes and compete along that dimension as well. Such *auction contests* can in fact be justified, as a contest that allows suppliers to bid on their reward is optimal (see Che and Gale 2003). Suppose the buyer in the above example lets suppliers bid prices for their innovations and then procures from the one offering the highest net surplus (the difference between the quality offered and payment demanded). This auction contest induces each supplier to randomize over q uniformly from [0,1] and to bid $\frac{1}{2}q$ for his prize, yielding a net expected surplus of $\frac{1}{3}\left(> \frac{8}{25}\right)$ to the buyer.

The purpose of employing competition here is not to select a supplier efficiently (recall both suppliers are equally efficient *ex ante*) but rather to provide incentives for unverifiable quality. The buyer has an incentive to select the supplier that offers the best value (quality minus payment), and this option to select from suppliers – just like the option to reject in the option contract – creates incentives for quality from the suppliers, and assures the surplus for the buyer even without supplier buy-in. Such incentives come at the expense of duplicative investment, however, since the buyer procures from only one supplier. This suggests that limiting the number of competitors – often to two – is optimal (see Che and Gale 2003; Taylor 1995; Fullerton and McAfee 1999). If the quality of the innovation/good is stochastic, competition will serve the additional purpose of identifying an efficient supplier.

The non-verifiable quality problem can be overcome if the buyer procures the good repeatedly. In such a situation, a *reputation* – more specifically, the promise of granting rents in exchange for an agreed-upon quality – combined with the threat of terminating a relationship for a sub-par quality, can create the supplier's incentives for quality (Klein and Leffler 1981). For instance, in the above example the buyer and a supplier can make an implicit agreement such that the latter provides the quality of $q^* = 1$ in exchange for a payment $p \in \left(\frac{1}{2}, 1\right)$ from the former, as long as both honour the agreement; if one deviates unilaterally, both terminate the relationship. The threat of termination is credible, since it is a Nash equilibrium in each round for the supplier to provide zero quality and for the buyer to pay nothing. If nobody deviates, the buyer and the supplier would obtain $\frac{1-p}{1-\delta}$ and $\frac{p-0.5}{1-\delta}$ respectively, where $\delta \in [0, 1)$ is a common discount factor. The two parties can get at most 1 and p, respectively, from a unilateral deviation. If $\delta \geq \left\{p, \frac{1}{2p}\right\}$, then the first-best quality can be implemented in a subgame perfect equilibrium.

So far, we have assumed that quality of procurement is observable to the buyer (albeit non-verifiable to the court). Often, the quality of good supplied may not be observable to the buyer at the time she procures from a supplier. Development of new weapons or transportation systems are subject to this problem, as the quality of new features is learned long after the procurement. If the buyer is unsure about the quality supplied, a standard auction based solely on price performs poorly (and unobservability of quality precludes the use of multi-dimensional competition such as scoring auctions). In such a case, it may be socially optimal for the buyer to bargain with one of many potential suppliers, instead of inviting them to compete for a job. To illustrate, suppose there are two potential suppliers, each with cost c drawn independently and uniformly from [0,1]. A supplier with c can deliver a good with quality $v(c) = 3c$ to the buyer, so the quality is not only unknown to the buyer but also positively correlated with the supplier's cost. In this case, competition based only on price will result in the selection of the low quality, with the buyer obtaining only $\frac{1}{3}$ in expectation. By contrast, if the buyer selects a supplier at random and makes a *take-it-or-leave-it offer* of price 1, then the offer will be accepted, and the buyer will enjoy an expected surplus of $\frac{1}{2}$. Notice that bargaining dominates competition also in social surplus in this case (see Manelli and Vincent 1995).

Procurement Irregularities

The difficulty with verifying quality may require a buyer to hire agents with special expertise to evaluate the proposals. This added bureaucracy can introduce agency costs to the procurement. In particular, there is a potential for the agents evaluating proposals to favour a certain supplier in exchange for a bribe or kickback. *Corruption* is a serious problem in both public and private procurement, particularly across national borders. (Between 1994 and 1999, bribery was allegedly a factor in the awarding of nearly 300 contracts worldwide worth $145 billion and caused US firms to lose as many as 77 contracts worth $24 billion.) Burguet and Che (2004) analyse this problem via a scoring-auction model where quality score is measured imperfectly and is manipulable by the procurement agent in exchange for a bribe, and show that bribery competition – unlike standard auction competition – leads to allocational inefficiencies (see also Celentani and Ganuza 2002; Burguet and Perry 2002; Compte et al. 2005).

Another type of procurement irregularity is *collusion* among bidders in procurement competition. Bidding cartels in procurement auctions account for a significant portion of antitrust cases. McAfee and McMillan (1992) show that standard auctions are vulnerable to collusion but that the outcome will depend crucially on whether the cartel can exchange transfers. If the cartel members can exchange transfers, they can organize a ‘knock-out’ auction to achieve an efficient allocation, whereas if transfers cannot be used (for fear of detection, say) a member will be chosen randomly to win without any competition, meaning that allocation will be inefficient. Subsequent work has shown that repeated interaction allows asymmetrically informed cartel members to sustain collusion via a ‘bid rotation’-type scheme, and that the scheme can be refined to attain a degree of allocational efficiency (see, for example, Aoyagi 2003; Athey and Bagwell 2001; Athey et al. 2004; Blume and Heidhues 2002; Skrzypacz and Hopenhayn 2004). To what extent inefficiencies result from procurement irregularities and how they can be remedied by procurement policies remain open questions. (For some promising lead for the latter question, see Che and Kim 2006a; 2006b; Dequiedt 2005; Marshall and Marx 2003; Pavlov 2006).

See Also

- ▶ Auctions (applications)
- ▶ Auctions (theory)
- ▶ Cartels
- ▶ Defence Economics
- ▶ Hold-Up Problem
- ▶ Incomplete Contracts
- ▶ Mechanism Design
- ▶ Mechanism Design (New Developments)

Bibliography

Aoyagi, M. 2003. Bid rotation and collusion in repeated auctions. *Journal of Economic Theory* 112: 79–105.

Asker, J. and Cantillon, E. 2004. *Properties of scoring auctions*. Discussion Paper No. 4734. London: CEPR.

Asker, J. and Cantillon, E. 2005. *Optimal procurement when both price and quality matter.* Discussion Paper No. 5276. London: CEPR.

Athey, S., and K. Bagwell. 2001. Optimal collusion with private information. *The RAND Journal of Economics* 32: 428–465.

Athey, S., K. Bagwell, and C. Sanchirico. 2004. Collusion and price rigidity. *Review of Economic Studies* 71: 317–349.

Bajari, P., and S. Tadelis. 2001. Incentives versus transaction costs: A theory of procurement contracts. *The RAND Journal of Economics* 32: 387–407.

Bajari, P., McMillan, R. and Tadelis, S. 2002. *Auctions versus negotiations in procurement: An empirical analysis*. Working paper, Department of Economics, Stanford University.

Baron, D., and R. Myerson. 1982. Regulating a monopolist with unknown costs. *Econometrica* 50: 911–930.

Blume, A., and P. Heidhues. 2002. *Modeling tacit collusion in auctions*. Mimeo: University of Pittsburgh.

Burguet, R., and Y.-K. Che. 2004. Competitive procurement with corruption. *The RAND Journal of Economics* 35: 50–68.

Burguet, R. and Perry, M. 2002. *Bribery and favoritism by auctioneers in sealed bid auctions*. Working paper, Department of Economics, Rutgers University.

Celentani, M., and J.-J. Ganuza. 2002. Corruption and competition in procurement. *European Economic Review* 46: 1273–1303.

Che, Y.-K. 1993. Design competition through multidimensional auctions. *The RAND Journal of Economics* 24: 668–680.

Che, Y.-K., and I. Gale. 2003. Optimal design of research contests. *The American Economic Review* 93: 646–671.

Che, Y.-K., and D. Hausch. 1999. Cooperative investments and the value of contracting. *The American Economic Review* 89: 125–147.

Che, Y.-K. and Kim, J. 2006a. Robustly collusion-proof implementation. *Econometrica* 74, 1063–108.

Che, Y.-K. and Kim, J. 2006b. *Optimal collusion-proof auctions*. Discussion Paper No 0506–22, Department of Economics, Columbia University.

Compte, O., A. Lambert, and T. Verdier. 2005. Corruption and competition in procurement auctions. *The RAND Journal of Economics* 36: 1–15.

Corts, K., and J. Singh. 2004. The effects of relationships on contract choice: evidence from offshore drilling. *Journal of Law, Economics, and Organization* 20: 230–260.

Crocker, K., and K. Reynolds. 1993. The efficiency of incomplete contracts: an empirical analysis of Air Force engine procurement. *The RAND Journal of Economics* 24: 126–146.

Dequiedt, V. 2005. *Optimal collusion and optimal auctions*. Mimeo: University of Toulouse.

Fullerton, R., and R. McAfee. 1999. Auctioning entry into tournaments. *The Journal of Political Economy* 107: 573–605.

Kahn, C., and G. Huberman. 1988. Two-sided uncertainty and up-or-out contracts. *Journal of Labor Economics* 6: 423–443.

Klein, B., and K. Leffler. 1981. The role of market forces in assuring contractual performance. *The Journal of Political Economy* 89: 615–641.

Laffont, J.-J., and J. Tirole. 1986. Using cost observations to regulate firms. *The Journal of Political Economy* 94: 614–641.

Laffont, J.-J., and J. Tirole. 1987. Auctioning incentive contracts. *The Journal of Political Economy* 95: 921–937.

Manelli, A., and D. Vincent. 1995. Optimal procurement mechanisms. *Econometrica* 24: 668–680.

Marshall, R., and L. Marx. 2003. *Bidder collusion*. Mimeo: Penn State University.

McAfee, R., and J. McMillan. 1986. Bidding for contracts: A principal agent analysis. *The RAND Journal of Economics* 17: 326–338.

McAfee, R., and J. McMillan. 1987. Competition for agency contracts. *The RAND Journal of Economics* 18: 296–307.

McAfee, R., and J. McMillan. 1992. Bidding rings. *The American Economic Review* 82: 579–599.

Myerson, R. 1981. Optimal auction design. *Mathematics of Operations Research* 6: 58–73.

Pavlov, G. 2006. *Colluding on participation decisions*. Working paper, Department of Economics, Boston University.

Riordan, M., and D. Sappington. 1987. Awarding monopoly franchises. *The American Economic Review* 77: 375–387.

Skrzypacz, A., and H. Hopenhayn. 2004. Tacit collusion in repeated auctions. *Journal of Economic Theory* 114: 153–169.

Taylor, C. 1993. Delivery-contingent contracts for research. *Journal of Law, Economics, and Organization* 9: 188–203.

Taylor, C. 1995. Digging for golden carrots. *The American Economic Review* 85: 872–890.

Producers' Markets

Harrison C. White and Robert G. Eccles

Any market is a social formation which decouples sellers from buyers exactly by turning the particular persons into occupants of roles. These roles form a transposable structure, which also translates items of offer into roles as commodities. Other varieties of such social formations are, for example, ritual prestation cycles of gifts, in which status and purity are computed via regularized offerings and receptions (such as Strathern's *The Rope of Moka*, 1971). But all markets are decentralizing; they dissolve the global structure of flows in prestation institutions into locally accountable flows.

The product market is by any account the predominant modern form. It is peculiarly concerned with asymmetry. Producers' markets generate and guarantee continuing flows of production from dedicated producers, but they exhibit flexibility and variety with respect to the sorts of buyers and their organizational forms. Producers' markets evolved historically from *verlager, kaufman*, and putting-out systems of early modern times (Kriedte et al. 1981) in which production became increasingly rationalized and concentrated within more and more formal organization. Interchangeable sellers and buyers of markets became specialized into producer/seller and buyer.

Production Markets as Role Structures

Asymmetry is problematic for role structures, because asymmetry can only be maintained between roles when they are explicitly articulated into a global structure such as the Rope of Moka. Role structures are more flexible and effective and translatable when they define roles as parallel, as structurally equivalent in relation to the ongoing processes (White et al. 1976). Competition is an important example. Those who compete are *a fortiori* similar; they can be compared and are comparable one to another.

A producers' market organizes producers into an array of parallel roles whose primary focus is each other. The producers' market exists to remove the need for attention to individual roles on the 'other side' of the market from producers who develop roles with respect to each other. In common business parlance, these roles are articulated as the producer's *strategy*, which details the role the producer hopes to occupy in the market in terms of volume, cost structure, and quality. It is revealing to recognize that these expressions of strategy, which can as easily be made *post hoc* as *a priori*, are rarely made solely in absolute terms. Instead they are made with respect to other producers with whom one is competing. These others are, in turn, attempting to do the same thing and so a certain recursive element is introduced. Each producer defines its role in terms of the similarities and differences it has with respect to other producers.

Markets vary in the accuracy of their producers' ability to make these comparisons. In some markets these perceptions fairly closely reflect underlying economic realities and buyers' perceptions. This contributes to the reproducibility of the social structure. When producers' perceptions are inaccurate, such as when each producer thinks of itself as different from all the others each of whom thinks it is different from all the others in exactly the same way, roles are ambiguous and the resulting market is unstable.

The substance of a producers' market as a social formation is the frame or schedule for terms-of-trade. These are perceived as objectively binding by the actors deciding within them; yet their basis is exactly the choices just made by producers. A producers' market is a tangible social construction, an interface which can only be maintained through the reproduction of its own assumptions in the very course of acting in terms of it. This requires accurate perceptions and the willingness of each producer to act in the future similarly to how it has acted in the past or, at a minimum, to change its actions in predictable ways. Nothing wreaks as much havoc in a producers' market as unanticipated actions by one or more producers. A particularly salient example is producers' prices. Businessmen often comment that nothing is worse than a competitor who prices 'irrationally', i.e., unpredictably. Thus, while producers' roles are defined in terms of their competitive relationships with each other, the existence of a market requires some form of tacit cooperation to continue to play by the established rules of the game.

There are two fixed sides and there also are two sorts of flows, physical volumes and value transfers, in a producers' market. Producers are the shapers of deals, the actors who offer terms of trade. Thus they can attempt to optimize outcomes for themselves, within the objective constraints of the terms of trade. In producers' markets buyers are sayers of yea/or/nay. Buyers refuse any deal which is not as good as other deals. In that way buyers enforce discipline on the terms of trade.

Producers' markets as social mechanisms presume and require small numbers of producers: the social mechanisms cannot work otherwise. Since role definition depends upon interproducer comparisons, the complexity of making these comparisons grows geometrically as the number of producers grows arithmetically. There are obvious limitations to how many producers any given producer can be sufficiently knowledgeable about in order to define its role with respect to these others. Consequently, there are limitations to how many actors can sustain the kinds of competitive and cooperative efforts that are necessary in order for a market to exist. A rough rule of thumb is that stable markets cannot contain more than a dozen or so producers.

Terms-of-trade seen by any producer are the revenues received for the various volumes shipped. These are the observables. They are easily extrapolated into a schedule of how revenue may change with volume. Everyone knows that producers differ in qualities and in cost structures, as well as in volumes shipped and revenues received, but only the latter are easily observable. Everyone also knows that buyers do discriminate among producers in ways summed up as quality, but no one can quantify this in advance or independent of volumes shipped. Instead producers recognize that responses to quality underlie and shape the observed schedule of terms-of-trade.

Evidence that the importance of quality differences to structuring the terms-of-trade is recognized by producers is seen from the expense and effort expended in market surveys, conducted by the producer itself or through retaining the services of a market research firm. An entire industry has emerged to help producers better understand how buyers perceive their quality and the quality of their competitors. This contributes to accurate role definitions (i.e. strategies) and thus to the reproducibility of the market. Of course, these surveys are not totally passive information gathering exercises. The execution of them contributes in a major way to the shaping of quality perceptions on the buyer side by defining the terms on which the elusive but very real notion of quality can be evaluated.

Similarly, an entire industry has emerged to provide information to producers on the cost structures of their competitors. The firms in what is often referred to as the 'strategy consulting industry' use a variety of information gathering mechanisms in order to estimate the cost structures of participants in a market. This information is obtained from both the producers themselves and from buyers. However, unlike the market research firms where the real value of the activity lies in shaping the buyers' perceptions of quality, information on cost structures is of relatively greater importance to producers. This is not to say that buyers are uninterested in the relationship between cost and price of their suppliers.

Pricing and Differentiation

Pricing is nonlinear with volume in production markets, reflecting differences in volumes shipped by producers of different quality. The price schedule tends to be monotonic and continuous, since buyers exert continuing discipline for deals of equal value, through which the schedule is extrapolated. The schedule will reproduce itself if each producer chooses again the production level of before. Given the perceived terms-of-trade, the producer will choose the volume at which revenue has the greatest margin over cost of production. An interesting consequence of the conditions for reproducibility is that producers must differ in profits and profitability in order to maintain the dispersion which sustains the market. While it may be true that eventually all profits become equal across firms and even across industries as capital seeks its best use, one of the strongest empirical findings of those who study actual markets is that profits and profitability differ substantially among them and between the firms within them. What actually *is* needs to be explained as much as what *will* (may) *be*.

A producers' market can be seen as a device for segmenting producers. Producer choices reflect off buyer gainsaying, which sorts out producers by volume in reflection of quality. A producers' market is not set up by – it does not come from – some pre-existing product. Rather, what a 'product' is emerges from successful evolution of producers' markets as self-reproducing schedules. The product defines the market and vice versa. Little is known of how new markets actually evolve – how often they are variants of pre-existing ones and how often they are new.

The driving force behind the evolution of markets is the effort of each producer to differentiate its offering from those of the other producers. In doing so it seeks to appeal to a specific segment of the buyer side. At the limit each producer would have its own set of customers which did not purchase from any other producer – markets of one firm. At the other extreme all buyers would purchase from any producer of a given product resulting in markets of dozens of producers. The creation of a self-reproducing social structure requires a market between these two extremes. This is achieved by a process that determines which producers will constitute the market, thereby defining the product, and which producers will be outside of the market and therefore members of a different one. Although this is difficult to determine analytically, it happens in business life all of the time.

Participants in a market know quite clearly who their competitors are (from the perspective of the producer) or who the available suppliers are (from the perspective of the customer). A product is defined in terms of the offerings made by the producers who constitute this market. The offerings of the producers in the market are considered more similar to each other than they are to offerings of producers not in the market.

Thus, markets require differentiation in two forms. The first is that which distinguishes market members from non-market members. The second distinguishes market members from each other in terms of their respective roles. As producers enter and leave the market, the product is constantly being redefined. One of the great competitive struggles in business life is between members of a market who seek to keep new entrants out and these potential new members who want in. Member cooperation supersedes member competition vis à vis non-members.

A basic implication is that producers' markets have arbitrary levels of 'supply and demand'. A market schedule is labile in the sense that buyers can only discipline the shape of the schedule, not the absolute levels of revenue and volume. A producers' market sorts out producers with respect to one another, but that regulates only their relative performance, not the absolutes. The ranges of price and volume within which a given market can and does emerge depends on historical circumstance.

Another important market characteristic which depends on historical circumstance is the pricing conventions used. Pricing conventions are a significant determinant of overall profitability levels. These pricing conventions include the units for which prices are established, the extent and nature of quantity discounts, how bundled or unbundled the pricing is of related products or services, returns policy, and the conditions in which prices can be renegotiated.

Prices are not something that mysteriously emerge from 'the market'. They are part of the terms-of-trade and are socially constructed by the actors involved in the exchange.

On the Nature of Markets

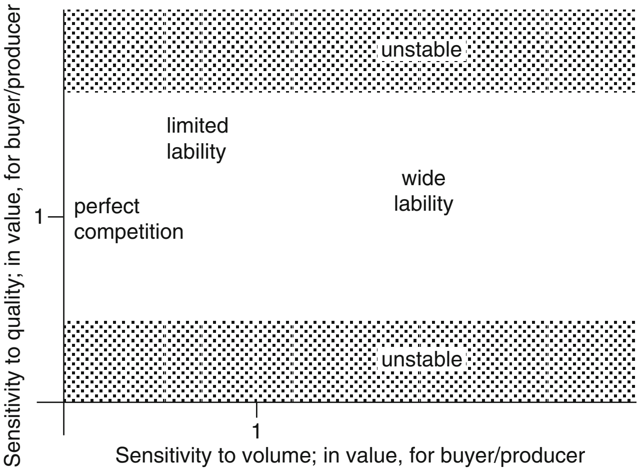
To sum up, a producers' market depends on and reflects the *dispersons* of producers and not their averages on costs and qualities. Producers necessarily need to differ in profits in order to reproduce the dispersion which sustains the market.

The market terms-of-trade is an interface which decouples producers from the buyer side. This permits the evolution observed into networks of markets with products of one being inputs to producers in many others as well as to ultimate consumers.

A major and heretofore unrecognized puzzle is resolved by this model of producers' markets as role structures: Why is it so rare to see two producers with equal shares in their market? Existing theories assume away the problem by hypothesizing a strict ordering of producers by market shares. Yet existing theory, on any plausible computation from the combinatorics of constituent firm characteristics, should expect to find equal shares a common phenomenon. The sociological view of production markets as role structures makes the missing equality possible.

Specific computations are necessary to bring this view of producers' markets to bear on specific current markets and to guide tests by comparison across a set of markets. Elsewhere (White 1981a,

Producers' Markets, Fig. 1 State space for production markets


unstable
limited
lability
wide
lability
perfect
competition
unstable
1
1
Sensitivity to quality; in value, for buyer/producer
Sensitivity to volume; in value, for buyer/producer

b, 1987; Leifer 1985; Leifer and White 1987) particular families of Cobb–Douglas functions have been used to frame such computations. By analytic continuity the main features of these results can be extended to less specific assumptions on cost and taste structure.

Figure 1 shows a state space consistent with the cited computations. The exact number and characteristics of producers are passed over. Instead two sensitivity ratios are used to identify possible states of a production market. The abscissa is sensitivity of taste to quality spread among producers, measured as a ratio, to the sensitivity of cost structures of producers to quality spread. This axis measures net differentiation among producers seen across the market. The ordinate is parallel, but sensitivities are to volume from a producer. This other axis can be seen as a measure of dependence which reflects tendency to concentrate purchase on a particular producer.

Existence is the crucial property to report: can producers' markets exist stably in that state? Since market schedules are labile, if a producers' market can exist for given sensitivity ratios it may exist with a range of particular terms-of-trade schedules. Detailed computations show that the answer depends in some states upon the range of producers included, but not in general. Only in one extreme limit is there no range of schedules possible and no limitation upon the spread of producers involved: this is the zero differentiation limit, on the left, corresponding to 'perfect competition'. Several results stand out:

1. Only when volume sensitivity is greater than quality sensitivity are robust markets common.
2. When volume sensitivity becomes large, markets collapse as production is concentrated in a single producer; yet there *is* a range of markets in states exhibiting some degree of increasing returns to scale.
3. In states yielding markets with increasing returns to scale, the addition of new producers to a market tends to decrease the total size of the market in gross revenue.
4. Dispersion in market share among producers tends to be less the more nearly equal the two sensitivity ratios are; as this equality is approached in the state space, markets are less stable and furthermore the dispersion in profitability declines, and average profitability increases.

This sociological perspective on markets is not without precedent in economics. It takes up again the realist stance opened up a half-century ago by Chamberlin but subsequently abandoned under the influence of P.A. Samuelson. Much of the technical apparatus underlying Fig. 1 is adapted from Spence's 1974 study of market 'signalling'. Various aspects are akin to themes in the economics literature, for example Sanderson (1974) on ‘demand’, Spulbur (1981) on nonlinear pricing, Telser (1981) on futures markets, and the economies-of-scope literature surveyed in Bailey and Friedlander (1982). But this role theory explains producers’ markets endogenously as social mechanisms rather than exogenously from boundary conditions.

See Also

- ▶ Marketplaces

Bibliography

Bailey, E.E., and A.F. Friedlander. 1982. Market structure and multiproduct industries. *Journal of Economic Literature* 20: 1024–1048.

Chamberlin, E.H. 1933. *The theory of monopolistic competition*. 8th ed. Cambridge, MA: Harvard University Press, 1963.

Kriedte, P., H. Medick, and J. Schlumbohm. 1981. *Industrialization before industrialization*. Trans. B. Schempp. London: Cambridge University Press.

Leifer, E.M. 1985. Markets as mechanisms: Using a role structure. *Social Forces* 64: 442–472.

Leifer, E.M., and H.C. White. 1987. A structural approach to markets. In *The structural analysis of business*, ed. M. Mizruchi and M. Schwartz. London: Cambridge University Press.

Sanderson, W.C. 1974. Does the theory of demand need the maximum principle? In *Nations and households in economic growth*, ed. P.A. David and M.W. Reder, 173–221. New York: Academic Press.

Spence, A.M. 1974. *Market signalling*. Cambridge, MA: Harvard University Press.

Spulbur, D.F. 1981. Spatial nonlinear pricing. *American Economic Review* 71: 921–933.

Strathern, A. 1971. *The rope of Moka*. London: Cambridge University Press.

Telser, L.G. 1981. Why are there organized futures markets? *Journal of Law and Economics* 24: 1–22.

White, H.C. 1981a. Where do markets come from? *American Journal of Sociology* 87: 517–547.

White, H.C. 1981b. Production markets as induced role structures. In *Sociological methodology 1981*, ed. S. Leinhardt, 1–57. San Francisco: Jossey-Bass.

White, H.C. 1987. Varieties of markets. In *Social structures: A network approach*, ed. B. Wellman and S.D. Berkowitz. London: Cambridge University Press.

White, H.C., S.A. Boorman, and R.L. Breiger. 1976. Social structure from multiple networks. *American Journal of Sociology* 81: 730–780.

Product Cycle

Raymond Vernon

No one will deny that products come into existence, change in character, and eventually disappear or become altered out of all recognition. Archaeologists, historians, businessmen and ordinary people have no difficulty in recognizing that fact. Economists, at least since David Hume, have occasionally acknowledged the phenomenon. But until a decade or two ago, the disposition of economists to use that process as a basis for formal inductive or deductive analysis has been extraordinarily limited. The ideas have surfaced from time to time in the works of John Williams (1947), Donald MacDougall (1957) and Michael Posner (1961) among many others, only to slip back into limbo. When economists have found it difficult to disregard innovation and product change in any formal analysis, they have usually assumed that the economic effects of innovation could be captured through its consequences for increased productivity, hence through a change in production costs. The appearance of the railroad, the automobile and the commercial aircraft, therefore, was usually thought to have been captured by referring to a decline in the cost of transportation; the appearance of nylon was treated as the equivalent of a decline in the cost of thread.

Meanwhile, businessmen and business schools came to speak of products as going through a life cycle. During the course of that cycle, predictable changes were thought to occur in product characteristics, hence in production technique and production location. Equally predictable changes were seen in the nature of the market during the life cycle of a product, including changes in the prevalence of competition and in the character of demand.

In their early stages, new products tended to be unstandardized in character, as suppliers experimented with different inputs, different production processes, and different designs of the final product. This was a stage, therefore, in which suppliers could not readily determine an optimum location, production scale or sale price, and in which product differentiation among suppliers was relatively high. In a later stage, the number of suppliers would increase, the product would become more standardized, production and location decisions would be made with less uncertainty, and the price elasticity of demand – both in the aggregate and for the output of the individual supplier – would increase. Thereafter, some producers might elect to attempt some degree of product differentiation, using trade names, advertising and minor variations in product; but price elasticities at the firm level would still be quite high, demanding close attention to production costs.

These ideas, which have formed the basis of a large literature about product life cycles or product cycles over the past few decades, would ordinarily be assigned by any economist who took note of them to the microeconomics of imperfect markets. In the two or three decades immediately following World War II, however, some economists professed to see such considerable strength in the product cycle factor as to help explain the macroeconomic performance of some countries, notably, the foreign trade patterns and foreign direct investment patterns of the US economy.

In brief, the product cycle was thought to be greatly influencing the performance of the US economy as a whole because of some highly distinctive features of that country. For many decades, the United States had recorded the highest per capita incomes in the world, along with high hourly labour costs, low capital costs and low costs of raw materials. That configuration, according to the argument, had placed a distinctive cast on US innovation. US innovators specialized in generating new products that responded to the emerging demands of a mass market of high-income consumers, as well as to the desires of producers for devices that could produce goods and services with less labour.

Fortuitously, according to the argument, the US pattern of innovations created a stream of products during much of the twentieth century that subsequently would be appropriate for other countries as well. That outcome was a result of the fact that during this period, the per capita incomes and labour costs of other countries were also rising, albeit from a lower level, while the capital costs and raw material costs of these countries were declining in relative terms; hence the products that had been generated for the US economy in 1950 became increasingly appropriate for other economies in the succeeding years.

These asserted relationships, supported by numerous empirical studies, were thought to provide an important part of the explanation for the strong showing of the US economy in the exportation of manufactured products, especially in high-technology goods, along with the heavy investments of US firms in subsidiaries in foreign markets devoted to the manufacture of such products.

By the 1970s, however, it was evident that the performance of US firms in both exports and direct investment was changing. At the same time, the product cycle factors that were thought to have produced the strong international performance of the US economy up to that time also were seen as considerably altered. For one thing, the per capita incomes of Western Europe and Japan were rapidly rising to the US level. Moreover, the distinctiveness in the profile of US factor costs was being obliterated, as the global costs of capital and raw materials came to be reflected increasingly within US markets. As a result, Europe-based and Japan-based innovators, responding mainly to the conditions of their own home markets, were found to be challenging US innovators with similar products.

Moreover, the 1970s interrupted the long-time global trends that had created markets for US innovations. Income growth was interrupted; capital costs and material costs rose more rapidly than labour costs. Now it was the turn of the European and Japanese innovators, with their long-time emphasis on the conservation of materials, the improvement of product performance, and the paring of costs. As a consequence, the United States found itself importing on a large scale products in which such characteristics had come to matter, such as steel, electronics and automobiles.

The power of the product cycle concept as an explicator of the export and investment behaviour of the US economy accordingly declined in the 1970s. As US products matured in world markets, US producers were pressed increasingly to concern themselves with the costs and prices of those products. Moreover, once the US firms had established their multinational networks, the existence of those networks gradually altered the perceptions and calculations of some US-based firms. Some began to respond to conditions in foreign markets, as well as the home market, to provide the stimuli for their innovations. Some began to think in terms of world models for their products, with production facilities for component parts established in any country in the world where factor cost considerations might indicate. Trends of this sort reduced the relevance of the nationality of the parent in determining the direction of the firm's innovations and in determining the patterns of its imports and exports.

Still, the product cycle concept continues to have some considerable utility. Most producing firms continue to produce within a single national economy, even though the relative output of such firms is considerably smaller than their relative numbers. Moreover, national economies continue to retain some distinctive national characteristics, a fact that may lead their entrepreneurs to generate distinctive new products. Thus, US and Israeli producers operate in economies that are heavy producers of military goods; Argentine and Indian producers in economies in which they are challenged to overcome the handicaps imposed by unreliable supplies of power or transportation; Singapore producers in a national economy so small that it cannot provide the basis for scale economies. These different conditions will tend to push innovations in somewhat different directions, creating fertile grounds for theorizing about the trade and investment patterns that these differences will eventually produce.

See Also

- ▶ Diffusion of Technology
- ▶ Innovation

Bibliography

MacDougall, Sir D. 1957. *The world dollar problem*. London: Macmillan.

Posner, M.V. 1961. International trade and technical change. *Oxford Economic Papers* 13: 323.

Williams, J.H. 1947. The theory of international trade reconsidered. Reprinted as ch. 2 of his *Postwar monetary plans and other essays*. Oxford: Blackwell.

Product Differentiation

Simon P. Anderson

Abstract

Product differentiation is pervasive in markets. It is at the heart of structural empiricism and it smoothes jagged behaviour that causes paradoxical outcomes in several theoretical models. Firms differentiate their products to avoid ruinous price competition. Representative consumer, discrete choice and location models are not necessarily inconsistent, but performance depends crucially on the degree of localization of competition. With (symmetric) global competition, rents are typically small and market variety near optimal. With local competition, profits may be protected because entrants must find profitable niches. These rents lead firms to competitively dissipative them, and performance may be poor.

Keywords

Bertrand competition; Bertrand paradox; Business stealing; Chain linking; Characteristics; Circle model; Constant elasticity of substitution (CES) model; Diamond paradox; Discrete choice models; Endogenous growth; General probit model; Horizontal and vertical differentiation; Intra-industry trade; Local competition; Location models; Market power; Menu costs; Monopolistic competition; Nested logit model; Network externalities; New Keynesian macroeconomics; Product differentiation;

Quality ladders; Representative consumer models; Spatial competition; Vertical differentiation

JEL Classifications
L13

Overview

Consumer goods are available in a variety of styles and brands. Product differentiation refers to such variations within a product class that (some) consumers view as imperfect substitutes. The store Foods of all Nations in Charlottesville, Virginia, USA (area population 120,000) carries 118 varieties of hot pepper sauce, 41 balsamic vinegars and 121 different olive oils (these figures include variations such as flavourings and different package sizes from the same manufacturer). There are 82 other retail grocers listed in the area. Charlottesville is served by 23 rated radio stations which differ by format choices (18 are commercially operated).

Product differentiation offers firms market power. This enables them to transcend the Bertrand paradox for pricing homogeneous products. In the Bertrand paradox, two or more firms sell goods that consumers perceive as identical, so goods are perfect substitutes. Assume that marginal costs are common and constant, and market demand has a finite price intercept. Then one good cannot carry a price premium over another while retaining positive sales. Any lowest price above marginal cost would then profitably be undercut. This logic impels us to marginal cost pricing as the only equilibrium under Bertrand competition.

Product differentiation resolves the paradox naturally. When products are imperfect substitutes, a price-cutting firm cannot take all its rivals' customers with an infinitesimally small price cut. This means that firms have some market power (due to the special features that distinguish them from their rivals' products); they can set prices without a completely elastic response by consumers. It also means that the product itself becomes a choice variable and firms differentiate to avoid the Bertrand outcome.

However, many models of product differentiation do not treat this choice explicitly, and instead assume a framework (representative consumer, discrete choice, and symmetric location models) that generates a demand system. It is not so much the framework used but rather the structure of product differentiation that is critical to the predictions and results. Indeed, common models of one type may be recast within another framework and be formally equivalent. Instead, the important feature for performance is whether each product is equally substitutable with all others or if each has only few close substitutes which are chain-linked to other products in the industry. Equal substitutability describes *global competition* where each firm competes with each other firm. Chain-linking corresponds to *local competition*. Local competition models naturally apply in geographical space since nearby stores are closer substitutes for consumers than distant ones. Likewise, in a characteristics setting, a consumer with a sweet tooth will find sugary products closer substitutes for any sweet product than for a saltier one.

The next section describes models of product location (in geographical space or its characteristics analogue) and distinguishes horizontal from vertical differentiation. Section "Modelling (Horizontal) Product Differentiation" compares the common approaches to product differentiation used to analyse the market provision of variety. In these models, product decisions are suppressed and product selection is determined by entry. Section "Monopolistic Competition and Optimal Variety" describes how the market variety diverges from the equilibrium one. Section "Localized Competition" elaborates on this theme for local competition. Section "Further Applications" indicates how product differentiation is used elsewhere in economics.

Product Choice

Hotelling (1929) wrote the seminal paper treating the product specification as endogenous.

Applications beyond industrial organization include marketing, economic geography (spatial competition), political science (the 'Hotelling–Downs' model), and media economics. The basic paradigm is that consumers are differentiated by their locations ('addresses') and dislike distance. Products, too, are locations in this space (geographic, characteristics and so on). When products are priced at marginal cost, consumers differ by which they like best, a situation known as *horizontal differentiation.* The simplest version of the model has two ice-cream sellers locating on a beach (with fixed prices). The Nash equilibrium is back-to-back pairing at the median of the consumer distribution, a result christened the *principle of minimum differentiation.* It has been used to explain striking similarities in colas, petrol station location, political parties' platforms and the timing of television programmes.

However, the principle dissolves when firms locate in rational expectation of ensuing equilibrium prices (that is, seeking a subgame perfect equilibrium to a two-stage price-then-location game). Indeed, if two products were collocated, Bertrand competition would drive prices to marginal cost. Firms will avoid this ruinous result by differentiating to retain market power attributable to location advantage. The equilibrium trades off two opposing factors. Getting closer to a rival provokes more intense price competition, so firms differentiate in order to relax price competition, but getting close to a rival attracts more customers.

The equilibrium locations are outside the optimum ones (which are at the quartiles for a uniform consumer density) for the central case of quadratic distance disutility costs, but otherwise there is no fundamental reason for excessive market differentiation. More elaborate models can rapidly become quite intractable and are hamstrung by non-existence of (pure-strategy) price equilibria due to fundamental failure of quasi-concavity of the profit functions in prices.

The case above of horizontal differentiation has consumers with fundamental preference differences across different varieties. In *vertical differentiation,* all consumers have the same preference ordering (when goods are priced at marginal cost). More preferred goods are often described as having higher 'quality' (with different individuals having different willingness to pay for quality). In vertical differentiation models, firms are to choose their product qualities. Choosing the same quality is avoided because of ruinous price competition, and the same trade-off operates as under horizontal differentiation. Under vertical differentiation though, the firm producing a higher-quality product earns more profit than a firm with lower quality. This result is an extension of the Bertrand paradox. One firm differentiates itself by a low quality, but this puts it at a disadvantage. Indeed, it may not be able to escape the shadow of the high-quality firm and earn a positive profit in equilibrium. This result implies the finiteness property that only a finite number of firms can survive in equilibrium even as fixed costs become arbitrarily small. By contrast, in a horizontal model, a firm may always find a niche between existing firms that gives it an advantage over some consumers (so that finiteness cannot hold). Finally, if the costs of improving quality are mainly sunk, a firm may invest more heavily in quality in a larger market because the benefit accrues over a larger consumer base (so sunk costs are endogenous).

Quite similar in spirit to the above approaches, Lancaster's (1966) model of *characteristics* was a quite revolutionary approach to consumer theory. It posited that consumers care about the characteristics intrinsic to goods and purchase goods because they deliver the desired characteristic mix, adjusting appropriately for prices. Lancaster's theory answers the question of why goods are desirable by formulating fundamental preferences over characteristics. The approach is intuitively appealing and is at the heart of hedonic models in econometrics, state preference and mean variance models in portfolio choice problems in finance, and structural econometric work in industrial organization. However, the approach is rather cumbersome for generating much theoretical insight into firms' location decisions, that is, the choice of which characteristics to embody in products.

Modelling (Horizontal) Product Differentiation

There are three basic families of product differentiation models that are typically used for modelling equilibrium with free entry and comparing optimal to equilibrium diversity.

Representative consumer models start by positing a utility function intended to portray aggregate preferences. This preference ordering generates the demand system for differentiated products and it measures welfare for the optimality analysis. Such functions typically embody global competition insofar as demands for varieties of the differentiated product are symmetric substitutes. Models in this class include the often-used constant elasticity of substitution (CES) preference formulation and the quadratic utility that gives rise to a linear demand system. These are parameterized utility functional forms that embody taste for variety in that more variety raises welfare even when total consumption is fixed.

The *discrete choice approach* is founded in econometric and probabilistic models of consumer behaviour. Each individual has an idiosyncratic taste (or ‘match value’) for each product. Aggregating individual choices yields the demand function and aggregating the surpluses yields the welfare function. Any i.i.d. tastes yield global competition in that products are symmetric substitutes (for example, the logit model).

Discrete choice models are not constrained to symmetric substitutability among variants. Models such as the nested logit embody closer substitutability between products within the same nest and the general probit model embodies quite elaborate substitutability patterns through the variance–covariance matrix of the match terms. These models are commonly used in the new structural empirical industrial organization literature.

Location models explicitly describe product specifications and consumer preferences as addresses and assume that consumers dislike distance ‘travelled’ between ideal type and product. Location models may also be viewed as discrete choice models because individuals make discrete choices and have idiosyncratic match values. There is a difference in interpretation: location models typically assume the population of consumers to be given and deterministic, while discrete choice models suppose that an individual’s taste is a realization from a probability distribution.

In models such as the circle model, the emphasis is on the number of products produced in equilibrium and exogenous symmetric locations are effectively imposed: however, the standard symmetric location pattern can be proved to be a location equilibrium under some circumstances.

One major benefit of discrete choice and location models is that the explicit micro foundations indicate how to introduce some economic phenomenon of interest. For example, network externalities may be incorporated into consumer utility and a consistent set of demands is then generated. Representative consumer models are less satisfactory since they do not start with a population of differentiated individuals.

The different approaches are not necessarily inconsistent with or substitutes for each other. Rather, they may frequently be twinned and one approach may be reinterpreted within the setting of the others. The CES model is a variant of the logit model, and a representative consumer exists for the circle model and for probabilistic discrete choice models. Indeed, although global competition is typically generated from models such as the CES representative consumer or models of discrete choice with i.i.d. errors, it can also be derived from a spatial model if there are sufficiently many dimensions (so that each good can be a ‘neighbour’ to each other).

These models are also useful for comparative static analysis of changing patterns in industries in response to structural changes in cost structures, population growth, transport costs and consumer tastes. These descriptions are useful for urban economics, industrial organization, international trade, and economic geography.

Monopolistic Competition and Optimal Variety

In Chamberlin’s (1933) monopolistic competition model, products have downward sloping individual demands, yet there are so many firms that a free entry condition reasonably applies. With increasing returns to scale in production, there is a social trade-off between the benefits from variety and the costs of producing further varieties. The market equilibrium roughly embodies the same type of trade-off in so far as more firms enter if fixed costs are lower. Chamberlin concluded (although without explicit analysis) that the market equilibrium would reach ‘a sort of ideal’.

Under symmetric global competition, each entrant carves out its market share equally from established firms. Then, the number is the largest whole number at which profits are positive. This number of firms is tied down uniquely and zero profit is a reasonable approximation. Strategic behaviour by firms is scarcely relevant since there are virtually no profits to be had.

Later work showed Chamberlin to be right that the market would settle on the same amount of product diversity as the (zero-profit constrained) optimum in the central case of CES preferences (and for the logit model). Other discrete choice models lead to over-entry: this is exacerbated with asymmetric product qualities. The market may also bias against products with high fixed costs and inelastic demands. Multi-product firms choose inefficiently narrow product ranges in order to relax price competition: this effect exacerbates excessive entry of firms.

Although the symmetric CES/logit results (asymmetries aside) suggest that product differentiation is not much cause for performance concern, the alternative framework of the circle model typically generates substantial over-entry of firms.

The divergence between equilibrium and optimum product variety depends on the balance between two opposing forces. When a firm chooses to enter, it does not consider that its entry will benefit consumers. This *non-appropriation of consumer surplus* is therefore a positive externality that the firm does not internalize insofar that it cannot capture this surplus in its revenue. This force favours insufficient entry into the marketplace. It is the only force governing a multiproduct monopolist’s choice of how many products to introduce, so it provides too few products. However, in a competitive setting, a firm’s entry can also reduce the profits of existing firms. This *business stealing* is also not accounted for by the firm in its entry calculus because other firms’ profits do not affect its own bottom line. This negative externality encourages too many firms in equilibrium. For the CES model, these two forces exactly cancel out. For the circle model of localized competition, business stealing dominates and so there are too many firms. Loosely, prices fall quite slowly with entry in the circle model, meaning that too many firms are attracted.

Localized Competition

Vickrey (1964) can be credited with developing several important themes of spatial competition. He formulated the circle model, finding over-entry at the equilibrium, and noting that there may be multiple equilibria under localized competition because a new entrant must fit in a niche between existing firms. An entrant’s expected market space is substantially smaller than an incumbent’s. This effect is exacerbated because entrants rationally expect incumbent firms will react to new entry (in a new Bertrand–Nash price equilibrium) by cutting prices. Incumbents may earn substantially higher gross profits than the cost of entry that would be incurred by an entrant. There are then multiple equilibria. These range from the tightest packing at which incumbents just earn zero profits (and so are not induced to exit), to a loosest packing at which incumbents earn substantial profits (and entrants will not wish to set up).

The normative economics are very sensitive to the particular equilibrium selected. Typically, the equilibrium where the incumbents just make zero profits involves too many firms, while the loosest packing equilibrium involves too few firms. It is therefore crucial to determine which equilibrium is the reasonable description. The possibility of positive profits is also very important for market conduct because firms will strive to capture the rents attributable to advantageous locations. The deadweight losses due to rent seeking should be added to any inefficiency in location choice per se. Firms may commit capital early to a market that is growing over time in order to stake claim to locations that will later be profitable. Such capacity may be sunk before it is economically viable in terms of flow profit. The equilibrium locations are those of minimum packing (maximal spacing). However, a subsidy to encourage more entry might simply raise the amount of rents that are dissipated.

Thus, while performance under global competition may not generate much cause for concern, there may be substantial welfare losses in situations characterized by a strong degree of localized competition.

Further Applications

Product differentiation explains and resolves some other paradoxical results that obtain when products are assumed to be perfect substitutes. The *Diamond paradox* holds that the monopoly price prevails in the presence of small search costs even with many firms. Suppose consumers expected the monopoly price to be charged everywhere. Any firm pricing lower can raise its price and not lose consumers: a lower price attracts no consumers from other firms because a lower price is not expected. Any (rationally expected) price below the monopoly one is not an equilibrium because a firm can raise its price by an amount up to the search cost without losing any consumer who encounters it first. There is thus a striking discontinuity between the Bertrand and Diamond paradoxes as the search costs go from zero to a small positive value. Product differentiation smoothes the transition by allowing the consumers to shop for attributes other than purely price. A consumer may indeed find the price she expected at the first store sampled but still search further if she expects to find a better match. Her continued searching effectively brings firms into competition with each other. Firms therefore reduce prices to retain consumers who search for better matches.

The existence of a (pure strategy) price equilibrium in the original Hotelling model can be restored (through restoring profit function quasiconcavity) if there is sufficient non-locational product differentiation (through idiosyncratic preferences for products). This mechanism can restore the principle of minimum differentiation in locations, even with endogenous prices.

The standard Bertrand–Edgeworth pricing problem treats capacity constraints and, with homogenous products, has only mixed strategy equilibria. With sufficient product heterogeneity, pure strategy equilibria re-emerge since the benefits from undercutting are reduced. Likewise, standard models of positive network externalities typically exhibit multiple equilibria or no pure strategy equilibria. Unique pure strategy equilibria result with enough differentiation of products.

In international trade, product differentiation explains the empirical paradox of intra-industry trade; much bilateral trade is in the same product class. Furthermore, product differentiation is a source of gains from trade (in addition to the traditional comparative advantage in production and factor difference reasons) because of access to larger markets supporting more variety. Endogenous growth theory relies on product differentiation (typically with CES preferences or ‘quality ladders’ based on vertical differentiation) to rationalize continued research and development of new varieties. It is also a predominant feature as an agglomerative force in recent models of new economic geography. In macroeconomics, product differentiation models have been used to introduce imperfect competition. This is useful for providing micro-underpinnings to New Keynesian analysis. For example, in conjunction with ‘menu costs’ of switching prices, it gives rise to real effects to monetary policy.

See Also

- ▶ Chamberlin, Edward Hastings (1899–1967)
- ▶ Oligopoly
- ▶ Spatial Economics
- ▶ Spence, A. Michael (Born 1943)
- ▶ Vickrey, William Spencer (1914–1996)

Bibliography

Anderson, S.P., A. de Palma, and J.-F. Thisse. 1992. *Discrete choice theory of product differentiation*. Cambridge, MA: MIT Press.

Archibald, G.C., B.C. Eaton, and R.G. Lipsey. 1986. Address models of value theory. In *New developments in the analysis of market structure*, ed. J.S. Stiglitz and G. Frank Mathewson. Cambridge, MA: MIT Press.

Berry, S., J. Levinsohn, and A. Pakes. 1995. Automobile prices in market equilibrium. *Econometrica* 63: 841–890.

Chamberlin, E. 1933. *The theory of monopolistic competition*. Cambridge, MA: Harvard University Press.

D'Aspremont, C., J.J. Gabszewicz, and J.-F. Thisse. 1979. On hotelling's 'stability in competition'. *Econometrica* 47: 1145–1150.

Hotelling, H. 1929. Stability in competition. *Economic Journal* 39: 41–57.

Lancaster, K.J. 1966. A new approach to consumer theory. *Journal of Political Economy* 74: 132–157.

Spence, A.M. 1976. Product selection, fixed costs and monopolistic competition. *Review of Economic Studies* 43: 217–235.

Vickrey, W.S. 1964. *Microstatics*. New York: Harcourt, Brace and World [The section 'Spatial competition and monopolistic competition' repr. in *International Journal of Industrial Organization* 17 (1999): 953–963].

Product Life Cycle

Steven Klepper

Abstract

The product life cycle connotes the idea that, comparable to humans and other organisms, new industries evolve through distinct and predictable stages. When industries are young, they are subject to high product innovation, rapid output growth, a build-up in the number of producers, and flux in firm market shares. As industries age, product innovation gives way to process innovation, output growth declines, the number of producers goes through a shakeout, and firm market shares stabilize. Evidence supporting this characterization is discussed and three alternative theoretical accounts of it are reviewed.

Keywords
Innovation; Market share; Product life cycle

JEL Classification
L10

The product life cycle (PLC) connotes the idea that industries evolve through distinct and predictable cycles, similar to the way humans and other organisms pass through distinct stages in their lives. Originally, the idea of a PLC was proposed in the marketing literature (Levitt 1965). Subsequently it became a rallying point for how a number of disciplines view the evolution of new industries, especially ones with rich opportunities for product and process innovation.

Typically, three stages of evolution are distinguished in the PLC (see Williamson 1975, pp. 215–16; Clark 1985; and Drew 1987, for prototypical depictions). In the first stage, uncertainty about user tastes and the means to satisfy them is high, product design is primitive, unspecialized machinery is used in production, and the volume of production is low. Many firms enter and compete on the basis of product innovation, offering different variants of the industry's product. In the next stage, output growth is high, the design of the product begins to stabilize, product innovation declines, specialized machinery is substituted for labour, and the production process becomes more refined. Entry slows and exit exceeds entry, leading to a shakeout of producers. In the final stage, the industry becomes mature. Output growth slows, product innovation becomes less significant, diversity in product offerings declines, firm market shares stabilize, entry remains low and the number of firms may continue to decline, and management, marketing and manufacturing techniques become more refined. The firms that are left in the market disproportionately tend to be those that entered early.

Industries

This conceptualization of the PLC has been heavily influenced by the history of the US automobile industry, which is summarized in Klepper and Simons (1997). Commercial production of automobiles in the United States began in 1895. By 1904 annual sales were only 22,800 cars, but from 1909 to 1919 the average annual growth in the number of automobiles sold was 25.8%, and well over a million cars were sold in 1919. Subsequently annual growth slowed to 11.5% through 1929 and then declined to lower levels after the Second World War and the recovery from the Great Depression. Entry initially was slow but then rose steadily through 1907, when it peaked at 82 firms. Two years later the number of firms peaked at 272. Subsequently entry declined precipitously, the number of firms steadily fell, and by 1941 only nine firms were left in the industry. Counts of product and process innovations indicate that product innovation peaked early in 1905 but process innovation increased steadily into the 1930s, with innovations such as the moving assembly line revolutionizing the production process. Firm market shares initially fluctuated greatly, but after 1910 the industry was dominated by Ford and General Motors. Based on data for 1895–1966 (Klepper 2002), early entry provided a decided advantage. Thirteen of the 219 entrants from 1895 to 1904 survived at least 30 years, as against 3 of the 271 entrants from 1905 to 1909 and none of the subsequent 275 entrants from 1910 through 1966.

The shakeout of producers in automobiles was extreme, but Klepper and Simons (1997) document similarly severe shakeouts in tyres, television receivers and penicillin. All experienced rapid initial output growth that subsided over time. All experienced considerable entry that eventually became negligible, after which the number of firms declined for many years. In all three products firm market shares stabilized over time and the long-term survival rate was decidedly greater for earlier entrants (Klepper 2002). The record of innovation in the three products is less well documented than in autos (Klepper and Simons 1997). Trends in product and process innovation in tyres were similar to autos. In televisions there were only two major product innovations, both of which occurred early, but labour productivity grew steadily, suggesting no decline over time in process innovation. In penicillin process preceded product innovation, but this was largely due to a government orchestrated war effort to reduce the cost of production of penicillin rather than market forces.

These three products were part of a larger sample of 46 new products studied by Gort and Klepper (1982) and later by Klepper and Graddy (1990), Agarwal and Gort (1996), and Agarwal (1998). The products were typically characterized by high initial growth in output that declined over time. Pronounced shakeouts were common in a majority of the 46 products. Agarwal and Gort (1996) found that early entrants had greater survival rates, although this was also true of very late entrants. No systematic evidence was compiled on product and process innovation, but Agarwal's (1998) findings suggest that products subject to greater technological change were more likely to experience shakeouts.

Theory

Numerous theories have been proposed to explain 'excessive entry' and shakeouts, but three stand out in their emphasis on technological change and thus their ability to address all aspects of the PLC.

Jovanovic and MacDonald (1994) develop a model of a new industry that is created by a major invention. Initially firms enter to develop the invention until expected profits are driven to zero. Subsequently another major invention occurs that opens up the possibility of an increase in the minimum efficient scale of production. It may induce immediate entry, but entrants are at a disadvantage relative to incumbents in developing the invention because of their lack of experience. Successful innovators expand their output to the new minimum efficient scale, pushing down the price of the product until non-innovators are forced to exit, which triggers a shakeout.

Utterback and Suarez (1993) envision that firms enter a new industry based on innovative designs for the industry's product. Eventually consumers and producers coalesce around a particular design for the industry's product that becomes a de facto product standard known as a dominant design. Product innovation is limited to incremental improvements in the dominant design, which makes entry more difficult. Process innovation increases because firms become less fearful that product innovation will make investments in the production process obsolete. Firms less successful at process innovation, which tend to be later entrants with less experience, exit. Coupled with the decline in entry, this gives rise to a shakeout.

Klepper (1996) develops a model of the evolution of a new industry in which firm growth is costly and firm size conditions the returns from process R&D. This imparts a competitive advantage to earlier entrants. Over time, entry and incumbent expansion causes industry output to rise and price to fall. Eventually this renders entry unprofitable and it ceases. Continued decreases in price compromise the profitability of the latest entrants, forcing them to exit, giving rise to a steady decline in the number of producers. As incumbents expand, they increase their expenditures on process R&D, causing a rise in process relative to product innovation. Firm price–cost margins also get compressed, which diminishes the incentives of firms to grow, causing firm market shares to become more stable and industry output growth to decline.

Testing of alternative accounts of the PLC is in its incipiency. Klepper and Simons (2005) found that in autos, tyres, televisions and penicillin, innovation was dominated by the leading firms and was a key determinant of firm survival, as predicted by all three theories. During the shakeouts in the four products, exit rates of early and late entrants generally did not converge, which is not consistent with the first two theories.

Variations

Not all technologically progressive industries follow the PLC (Klepper 1997). Notably, some industries have not experienced any sign of a shakeout after 35 years and show no sign of the decline in entry that is characteristic of shakeout industries. Two examples that were in Gort and Klepper's (1982) sample of 46 new products are styrene, which is a petrochemical, and lasers. In both industries firms have ended up specializing either vertically (styrene) or horizontally (lasers). In other industries, technological developments led to turnover in the leading firms, undermining the advantages of early entry. This occurred in the disk drive industry before it went through a shakeout (Christensen 1993). It also occurred in autos, tyres and televisions well after their shakeouts had begun, with long-time US leaders displaced by Japanese and European firms that capitalized on small cars, radial tyres and the use of semiconductors in televisions.

These examples make clear that the PLC is a composite that only describes the prototypical evolutionary path followed by new industries. It will be an ongoing challenge to document systematic departures from the PLC and to understand the forces that contribute to them. This process is just beginning.

See Also

- ▶ Competition and Selection
- ▶ Evolutionary Economics

Bibliography

Agarwal, R. 1998. Evolutionary trends of industry variables. *International Journal of Industrial Organization* 16: 511–525.

Agarwal, R., and M. Gort. 1996. The evolution of markets and entry, exit, and survival of firms. *Review of Economics and Statistics* 78: 489–498.

Christensen, C. 1993. The rigid disk drive industry: A history of commercial and technological turbulence. *Business History Review* 67: 531–588.

Clark, K. 1985. The interaction of design hierarchies and market concepts in technological evolution. *Research Policy* 14: 235–251.

Drew, P. 1987. Despite shakeout, imaging industry not doomed to being Greek tragedy. *Diagnostic Imaging*, November. 95–99.

Gort, M., and S. Klepper. 1982. Time paths in the diffusion of product innovations. *Economic Journal* 92: 630–653.

Jovanovic, B., and G. MacDonald. 1994. The life cycle of a competitive industry. *Journal of Political Economy* 102: 322–347.

Klepper, S. 1996. Entry, exit, growth, and innovation over the product life cycle. *American Economic Review* 86: 562–583.
Klepper, S. 1997. Industry life cycles. *Industrial and Corporate Change* 6: 145–181.
Klepper, S. 2002. Firm survival and the evolution of oligopoly. *RAND Journal of Economics* 33: 37–61.
Klepper, S., and E. Graddy. 1990. The evolution of new industries and the determinants of market structure. *RAND Journal of Economics* 21: 27–44.
Klepper, S., and K. Simons. 1997. Technological extinctions of industrial firms: An inquiry into their nature and causes. *Industrial and Corporate Change* 6: 379–460.
Klepper, S., and K. Simons. 2005. Industry shakeouts and technological change. *International Journal of Industrial Organization* 23: 23–43.
Levitt, T. 1965. Exploit the product life cycle. *Harvard Business Review* 88: 81–94.
Utterback, J., and F. Suarez. 1993. Innovation, competition, and industry structure. *Research Policy* 22: 1–21.
Williamson, O. 1975. *Markets and hierarchies: Analysis and antitrust implications*. New York: Free Press.

Production and Cost Functions

Melvyn A. Fuss

The traditional starting point of production theory is a set of physical technological possibilities, often represented by a production or transformation function. The development of the theory parallels the firm's objective (cost minimization or profit maximization) and leads to input demands (and output supplies in the case of profit maximization) constructed from an explicit consideration of the underlying technology (i.e. derived directly from the production function).

An alternative approach to production theory is to start directly from observed economic data generated by markets – supplies, demands, costs and profits. In this case economic theory can be formulated in terms of the causal economic behaviour which is presumed to hold (cost minimization, profit maximization, revenue maximization, etc.), without the intervening constructive steps required in the traditional theory. The main advantage of such an approach is tractability. A direct approach to modelling the behavioural responses embodied in demand and supply functions simplifies the development of basic theoretical relationships and permits the specification and estimation of more complex production processes than is typically possible using the traditional, production function-based approach to production theory. This latter advantage is responsible for the virtual explosion of econometric estimation studies of production structures which have appeared in the last fifteen years based on the specification of economic behavioural relationships such as cost and profit functions.

It might appear that an approach to production theory based on economic observables such as cost would be less fundamental than one based on direct physical relationships. However, the theory of production duality establishes that, conditional on the optimizing behavioural response, the two approaches are equally fundamental and provide the same information about the production process. Using duality, any required characteristic of the underlying physical relationship can be uncovered from the observed economic variables.

In this entry I will concentrate on the duality between cost and production, although an analogous conceptual framework can be developed linking profit and production. Section "The Theory of Cost and Production Duality" presents the basic duality theorems which link cost and production functions as alternative, equally fundamental descriptions of the production technology. Section "History" provides a history of the development of the cost function approach to production theory. In section "Applications of Cost Functions" I survey several of the applications of the cost function approach which have appeared in the economic literature. The final section offers some concluding remarks.

The Theory of Cost and Production Duality

The *cost function* $C(\mathbf{y}, \mathbf{p}, \mathbf{z})$ is the *minimum* cost of producing a vector of outputs $\mathbf{y} = (y_1, \dots, y_M)$, when the firm faces a vector of exogenous input prices $\mathbf{p} = (p_1, \ \ldots, p_N)$, conditional on a set of exogenous characteristics of the production process $\mathbf{z} = (z_1, \ \ldots, z_R)$. The vector **z** represents a set of variables (other than outputs) over which the cost minimizing production unit (e.g. a firm) does not optimize. Examples include an index of technological change (usually a time trend) or the level of a fixed factor of production in a Marshallian short-run analysis. If the firm produces a single output *y*, and **p, z** are taken as fixed constants, then $C(\mathbf{y}, \mathbf{p}, \mathbf{z})$ reduces to the cost curve $C(y)$, a basic concept in elementary economic analysis.

Define the *production possibility set L.* to be the set of feasible input–output combinations **(x, y)** conditional on **z**. The boundary of *L.* can be represented by the *transformation function* ($F(\mathbf{y}, \mathbf{x}, \mathbf{z})$, where efficient production implies $F(\mathbf{y}, \mathbf{x}, \mathbf{z}) = 0$. In the case of a single output *y, F* = 0 can be solved for the *production function* $y = f(\mathbf{x}, \mathbf{z})$. The *input requirement set* $V(\mathbf{y}, \mathbf{z})$ is the set of all input bundles which can produce **y**, given **z**. The input requirement set corresponds to the conventional notion of an isoquant, but is well-defined for single or multiple output technologies. From the point of view of duality theory, it is convenient to represent the technology by the input requirement set rather than the production possibility set.

Duality theory between cost and production is based on the fact that if *V* or *F* possess certain properties, then *C* will possess a corresponding set of properties and vice versa. Consider the following set of regularity conditions on *V* and *F.* Suppose, conditional on **Z, V** is a nonempty closed, convex set characterized by free disposal of **x** or **y** (monotonicity). Alternatively, suppose (again conditional on **z**), that $F(\mathbf{y}, \mathbf{x}, \mathbf{z})$ is a real-valued continuous function which is monotone (increasing in **y** and decreasing in **x**) and quasi-convex in **y** and **x**. In the case of a single output *y,* the assumed properties of *F* imply that the production function $f(\mathbf{x}, \mathbf{z})$ will be, conditional on **z**, a real-valued continuous function which is monotone (increasing in **x**) and quasi-concave in **x**.

The Shephard (1953)–Uzawa (1964) duality theorems state that if *V* or *F* possess the above respective properties, then there will exist a cost function $C(\mathbf{y}, \mathbf{p}, \mathbf{z})$ with the following properties (contained on **z**):

(1) *C* is a positive real-valued function defined for all positive prices **p** and all positive producible outputs. In addition, $C(0, \mathbf{p}, \mathbf{z}) = 0$.
(2) *C* is non-decreasing in outputs and factor prices.
(3) *C* is a continuous function in y and **p**.
(4) *C* is a concave function in **p**.
(5) *C* is a linear homogeneous function in **p**.

The properties of the cost function with respect to **z** can be determined once one has knowledge of the nature of **z**. For example, in the case where **z** is a vector of fixed inputs, so *C* is a short-run variable (or restricted) cost function, *C* will be non-increasing, continuous and convex in **Z**.

The main force of the duality theorems is contained in the implication that, if we can specify a cost function possessing the regularity conditions enumerated in (1–5), then that cost function is guaranteed to embody a technology with the regularity properties specified above for *V, F* or *f.* In the words of McFadden (1978), p. 4), the duality between the cost function and the underlying production possibilities ‘establishes the cost function as a ‘sufficient statistic’ for all economically relevant characteristics of the underlying technology’.

While the duality theorems do not require differentiability for their validity, differentiability is almost always assumed in applications. Hence from this point on I will assume that *F* is twice differentiable in **(y, z)**, *f* is twice differentiable in **x**, and *C* is twice differentiable in **(p, y)**. Given differentiability, the cost function possesses two additional important properties:

(6) A derivative property known as Shephard’s (1953) Lemma, i.e.

$\frac{\partial C}{\partial p_i} = x_i,$

the cost-minimizing demand for the *i*th input, $x_i(\mathbf{p}, \mathbf{y}, \mathbf{z})$.
(7) A symmetry property analogous to the symmetry of the Slutsky matrix of consumer demand theory, i.e.

$$\frac{\partial^2 C}{\partial p_i \partial p_j} = \frac{\partial^2 C}{\partial p_j \partial p_i} \text{ or } \frac{\partial x_i}{\partial p_j} = \frac{\partial x_j}{\partial p_i}.$$

Shephard's Lemma can be used to derive systems of cost-minimizing input demand functions from arbitrarily specified cost functions possessing properties (1–5). The symmetry property can provide a test of the underlying cost minimization assumption or can be used to reduce the number of parameters which must be estimated in an econometric application, thus preserving degrees of freedom.

History

While the cost curve has a long history in economics, the cost and production functions literature is of relatively recent origin. Development of the literature had to await the understanding that an underlying production technology could be equivalently represented in either commodity or price space. Probably the first systematic approach to cost and production duality appears in two papers by Hotelling (1932, 1935). Hotelling analysed the problem of minimizing consumer expenditure subject to a utility level constraint, which is mathematically equivalent to the problem of minimizing cost subject to an output level constraint. He recognized the existence of a cost (expenditure) function concave in prices, introduced the profit function, and derived for the profit function a derivative property (Hotelling's Lemma) analogous to Shephard's Lemma. Hotelling also derived the symmetry conditions for the profit function which are analogous to those discussed above for the cost function.

The properties of expenditure functions were developed further by Roy (1942) and McKenzie (1957). The cost function and its properties were discussed in Samuelson (1947). Many of the developments in the modern theory and applications of cost and production functions have their origin in Shephard's (1953) pioneering book. Shephard exploited the theory of convex sets to establish rigorously the duality between cost and production functions. He also anticipated a number of the subsequent practical uses of duality theory: (1) as an aid in comparative statics analysis, (2) in econometric studies of production when cost and price data are more easily obtained than input data, and (3) as an aid in the analysis and use of aggregation conditions.

Additional early contributions to the duality theorems between *f* (or *F*) and *C* include McFadden (1978), Hanoch (1970, 1978) and Diewert (1971). Duality theorems between *V* (or *L*) and *C* have been proven by, among others, Uzawa (1964), McFadden (1966, 1978), Shephard (1970) and Diewert (1971).

The usefulness of cost functions and other dual functional forms (e.g. profit functions) in econometric applications was not recognized until Nerlove (1963) employed the Cobb–Douglas cost function in a study of returns to scale in the supply of electricity. The major impetus to empirical applications of cost functions occurred when Diewert (1969a, b) realized that cost functions representing more general production processes than the production functions then available for empirical analysis (e.g. Cobb–Douglas, C.E.S.) could be specified and easily estimated. From that realization came the popular Generalized Leontief (Diewert 1969a, b, 1971) and Translog (Christensen et al. 1970; 1973) functional forms. Later additions to the class of cost functions amenable to econometric estimation include: the normalized quadratic (Lau 1976), the Box-Cox (or generalized translog) (Brown et al. 1979), Berndt and Khaled 1979), and the Generalized McFadden (Diewert and Wales 1987).

In recent years there has been an enormous number of econometric applications of cost functions; far too many to list. Early applications which illustrated the advantages of the cost function approach for empirical analysis include Diewert (1969a), Fuss (1970, 1977b), Parks (1971), Burgess (1974) and Berndt and Wood (1975).

The duality approach of cost and production functions has begun to appear in the textbook literature. Textbooks currently in use which contain this method of analysis include Baumol (1977) and Varian (1984).

Applications of Cost Functions

In this section I will discuss two of the main applications of the cost function dual approach – to comparative statics and the specification of the characteristics of the production technology.

The Cost Function Approach to Comparative Statics

Suppose the cost function $C(\mathbf{y}, \mathbf{p}, \mathbf{z})$ is twice differentiable with respect to $\mathbf{p}$. Then since C is concave, the matrix of second order partial derivatives $\nabla^2_{pp}C$ is negative semi-definite, where

$$\nabla^2_{pp}C = \begin{bmatrix} \frac{\partial^2 C}{\partial p_1^2}, \frac{\partial^2 C}{\partial p_1 \partial p_2}, \ldots, \frac{\partial^2 C}{\partial p_1 \partial p_N} \\ \frac{\partial^2 C}{\partial p_2 \partial p_1} \\ \vdots \\ \frac{\partial^2 C}{\partial p_N \partial p_1}, \frac{\partial^2 C}{\partial p_N \partial p_2}, \ldots, \frac{\partial^2 C}{\partial p_N^2} \end{bmatrix} \qquad (1)$$

$$= \begin{bmatrix} \frac{\partial x_1}{\partial p_1}, \frac{\partial x_1}{\partial p_2}, \ldots, \frac{\partial x_1}{\partial p_N} \\ \frac{\partial x_2}{\partial p_1} \\ \vdots \\ \frac{\partial x_N}{\partial p_1}, \frac{\partial x_N}{\partial p_2}, \ldots, \frac{\partial x_N}{\partial p_N} \end{bmatrix}$$

The derivatives are to be interpreted as being evaluated at a particular cost-minimizing input bundle $x^*(\mathbf{p}^*, \mathbf{y}^*, \mathbf{z}^*)$, and to represent a small perturbation from $\mathbf{p}^*$. In particular, from the concavity property we have

$$\frac{\partial x_i}{\partial p_i} \leq 0, i = 1, \ldots, N \qquad (2)$$

which is known as the fundamental law of demand (input demand curves cannot slope upwards). Using Euler's theorem and the linear homogeneity of C yields the following restrictions on the derivatives of the input demand functions:

$$\left[\nabla^2_{pp}C\right] \cdot \mathbf{p} = \left[\nabla_p \mathbf{x}\right] \cdot p = 0. \qquad (3)$$

Finally we have (as before) the symmetry property of the second order partial derivatives of C, which implies a corresponding symmetry property for the derivatives of the input demand functions:

$$\frac{\partial^2 C}{\partial p_i \partial p_j} = \frac{\partial x_i}{\partial p_j} = \frac{\partial^2 C}{\partial p_j \partial p_i} = \frac{\partial x_j}{\partial p_i}. \qquad (4)$$

Cost-minimizing input demand functions must satisfy the comparative statics properties (2, 3, 4). One of the advantages of the cost function approach to production theory is the simplicity by which comparative statics can be analysed, in contrast with the original derivations (Hicks 1946; Samuelson 1947) which require the manipulation of bordered Hessian determinants.

A second type of comparative statics analysis which is facilitated by the cost function approach is the Marshallian short run–long run, or temporary equilibrium, approach to production modelling. Suppose $\mathbf{z}$ is now a vector of quasi-fixed inputs, fixed in the short run but variable in the long run. Examples of elements of $\mathbf{z}$ are capital plant and equipment and overhead labour. The cost function $C(\mathbf{p}, \mathbf{y}, \mathbf{z}) = C^s(\mathbf{p}, \mathbf{y}, \mathbf{z})$ is now a variable, or restricted, cost function and is interpreted as the minimum variable cost (in the short run) of producing $\mathbf{y}$, conditional on the level of quasi-fixed factors $\mathbf{z}$. The short-run input demand functions are obtained by applying Shephard's Lemma to C^s (i.e. differentiating C^s with respect to the prices of the variable factors p^i). Comparative statics in the short run is accomplished by applying to C^s the analysis contained in section "Applications of Cost Functions". The relationship between the short- and long-run cost functions and costminimizing input demand functions is obtained by noting that in long-run equilibrium the shadow values of the quasi-fixed factors, $-\nabla_z C^S$, must be equal to the market prices of these factors $\mathbf{q}$, i.e.

$$-\frac{\partial C^s}{\partial z_i} = q_i, i = 1, \ldots, R \qquad (5)$$

since the total cost of producing **y**,

$$\text{Totalcost} = C^s(\mathbf{p}, \mathbf{y}, \mathbf{z}) + \mathbf{q} \cdot \mathbf{z} \tag{6}$$

will be minimized with respect to **z** in the long run. Solving (5) for the long-run cost-minimizing levels of the quasi-fixed factors $\mathbf{z} = \mathbf{z}(\mathbf{p}, \mathbf{q}, \mathbf{y})$ and substituting in (6) provides the link between the long-run cost function $C^L(\mathbf{p}, \mathbf{q}, \mathbf{y})$ and the short-run cost function $C^s(\mathbf{p}, \mathbf{y}, \mathbf{z})$:

$$C^L(\mathbf{p}, \mathbf{q}, \mathbf{y}) = C^s[\mathbf{p}, \mathbf{q}, \mathbf{z}(\mathbf{p}, \mathbf{q}, \mathbf{y})] + \mathbf{q} \cdot \mathbf{z}(\mathbf{p}, \mathbf{q}, \mathbf{y}) \tag{7}$$

Differentiation of both sides of (7) with respect to elements of **p** or **y** provides the relationship between long-run and short-run comparative statics. For example,

$$\begin{aligned} \frac{\partial C^L}{\partial y_i} &= \frac{\partial C^s}{\partial y_i} + \sum_j \frac{\partial C^s}{\partial z_j} \cdot \frac{\partial z_j}{\partial y_i} + \sum_j q_j \frac{\partial z_j}{\partial y_i} \\ &= \frac{\partial C^s}{\partial y_i} + \sum_j \left[\frac{\partial C^s}{\partial z_j} + q_j \right] \cdot \frac{\partial z_j}{\partial y_i}. \end{aligned} \tag{8}$$

Equation (8) provides the relationship between long-run and short-run marginal costs. If the firm is in long-run equilibrium, then (5) holds and (8) reduces to the well-known equilibrium result that long-run and short-run marginal costs are equal.

The short run-long run illustration of temporary equilibrium analysis provides another example of the usefulness of the cost function methodology. Berndt and Fuss (1986) provide additional examples of the cost function approach to issues in temporary equilibrium. The ability to specify the cost function directly leads to important simplifications in analysing complex production processes.

The Specification of the Characteristics of Production

Economic effects such as scale, distribution, substitutability and technical change can be quantified in terms of the production function or the cost function and the first and second order derivatives of the respective function. Consider an n input production function $y = f(\mathbf{x}, t)$, where $\mathbf{x} = (x_1, \ldots, x_n)$ is a vector of inputs and t is an index of disembodied technical change. The corresponding cost function is $C = C(\mathbf{p}, y, t)$ where $\mathbf{p} = (p_1, \ldots, p_n)$. Table 1 contains a summary of the economic effects in terms of the production or cost function. The table contains $(n + 2)(n + 3)/2$ distinct effects which characterize the usual comparative statics properties of a production technology at a point in the input–output or input price–output space. The production function

Production and Cost Functions, Table 1 Economic effects and their relation to the production or cost function

Economic effect	Production function formula	Cost function formula	Number of distinct effects
Output or cost level	$y = f(x)$	$C = C(p, y, t)$	1
Returns to scale	$\mu = (\sum x_i f_i)/f$	$\mu = \frac{C}{y}/C_y$	1
Distributive share	$s_i = x_i f_i / \sum_1^n x_i f_i$	$s_i = C_i \cdot \frac{p_i}{C}$	$n - 1$
Own 'price' elasticity	$\varepsilon_i = x_i f_{ii}/f_i$	$\varepsilon_i = p_i C_{ii}/C_i$	n
Elasticity of substitution	$\sigma_{ij} = \frac{\left[-f_{ii}/f_i^2 + 2\left(f_{ij}/f_i f_j\right) - f_{jj}/f_j^2\right]}{\left[1/x_i f_i + 1/x_j f_j\right]}$	$\sigma_{ij} = \frac{C \cdot C_{ij}}{C_i C_j}$	$\frac{n(n-1)}{2}$
Rate of technical change	$T = f_t\ .\ /f$	$T = \ - C_t/C$	1
Acceleration of technical change	$\dot{T} = (f_{tt}/f) - (f_t/f)^2$	$\dot{T} = (C_t/C)^2 - (C_{tt}/C)$	1
Bias of technical change			
(i) rate of change of marginal products	$\dot{m}_i/m_i = f_{it}/f_i$		n
(ii) rate of change of cost shares		$\frac{\dot{s}_i}{s_i} = C_{it}/C_i - C_i/C$	n

Source: adapted from Fuss et al. (1978)

and its first two derivatives in (**x**, t) also comprise (n + 2)(n + 3)/2 distinct quantities; thus the production function formulae can be inverted to determine the function value and the first and second derivatives at a point in terms of economic effects. Since the cost function has n + 2 arguments in contrast to the n + 1 arguments of the production function, it may appear to permit a larger number of distinct effects involving first and second order derivatives. However, the restrictions on comparative static effects (from section "Applications of Cost Functions")

$$\sum_{i=1}^{n} p_i C_{ij} = 0, j = 1, \ldots, n \tag{9}$$

along with the additional restrictions

$$\sum_{i=1}^{n} p_i C_{yi} = c_y \tag{10}$$

$$\sum_{i=1}^{n} p_i C_{ti} = c_t \tag{11}$$

provide (n + 3) restrictions on the cost function and its derivatives. Therefore the number of distinct cost function conditions is

$$\frac{(n+3)(n+4)}{2} - (n+3) = \frac{(n+2)(n+3)}{2},$$

identical to the case of the production function. As before, the cost function formulae can be inverted to determine the function value and the first and second derivatives at a point.

The term *flexible functional form* has been applied to a production or cost function which has sufficient parameters to reproduce the (n + 2)(n + 3)/2 comparative statics effects at a point without imposing restrictions across these effects (Diewert 1974). The minimal number of parameters necessary for a functional form to be 'flexible' is (n + 2)(n + 3)/2, if unrestricted reproduction of the economic effects in Table 1 is to occur. Lau (1974) noted that a Taylor's series expansion to the second order would satisfy this criterion. For this reason cost functions which can be interpreted as second order approximations to an arbitrary cost function have become widely used in the empirical literature to estimate economic effects (or characteristics of the production structure). For example, the popular translog cost function (Berndt and Wood 1975; Fuss 1977a) approximates the logarithm of C(**p**, y, t) as a second order (quadratic) approximation in the logarithms of (**p**, y, t):

$$\begin{aligned}\log C &= \alpha_0 + \sum_{i=1}^{n} \alpha_i \log p_i + \alpha_y \log y + \alpha_t \log t \\ &+ \frac{1}{2}\left[\sum_i \alpha_{ii} (\log p_i)^2 + \alpha_{yy} (\log y)^2 + \alpha_{tt} (\log t)^2\right] \\ &+ \sum_i \sum_{\substack{j \\ i \neq j}} \alpha_{ij} \log p_i \log p_j + \sum_i \alpha_{iy} \log p_i \log y \\ &+ \sum_i \alpha_{it} \log p_i \log t + \alpha_{yt} \log y \log t\end{aligned} \tag{12}$$

The parameters of the cost function are

$$\alpha_0, \alpha_i, \alpha_y, \alpha_t, \alpha_{yy}, \alpha_{tt}, \alpha_{ij}, \alpha_{iy}, \alpha_{it}, \alpha_{yt}; \\ -\frac{(n+3)(n+4)}{2}$$

in number. The linear homogeneity regularity condition required of a cost function implies the following constraints on the parameters:

$$\begin{aligned}&\sum_i \alpha_i = 1; \sum_j \alpha_{ij} = 0, i = 1, \ldots, n; \\ &\sum_i \alpha_{iy} = 0; \sum_i \alpha_{it} = 0.\end{aligned} \tag{13}$$

Restrictions (13) imply (n + 3) parameter constraints; so that the number of free parameters of the translog cost function is

$$\frac{(n+3)(n+4)}{2} - (n+3) = \frac{(n+2)(n+3)}{2},$$

the minimal number required of a flexible functional form. The translog cost function formulae corresponding to the economic effects of Table 1 are set out in Table 2. One important characteristic of the translog cost function which is readily

Production and Cost Functions, Table 2 Economic effects and the translog cost function

Economic effect	Translog cost function formula
Cost level	$logC = logC(\log\mathbf{p}, logy, logt)$ as in (12)
Returns to scale	$\mu = \left[\alpha_y + \alpha_{yy}\log y + \sum_i \alpha_{iy}\log p_i + \alpha_{yt}\log t\right]^{-1}$
Distributive share	$s_i = \alpha_i + \sum_j \alpha_{ij}\log p_j + \alpha_{iy}\log y + \alpha_{it}\log t$
Own price elasticity	$\varepsilon_i = \left(\alpha_{ii} + s_i^2 - s_i\right)/s_i$
Elasticity of substitution	$\sigma_{ij} = (\alpha_{ij} + s_i s_j)/s_i s_j$
Rate of technical change	$T = -\frac{1}{t}\left[\alpha_t + \alpha_{tt}\log t + \sum_i \alpha_{it}\log p_i + \alpha_{yt}\log y\right]$
Acceleration of technical change	$y = f(x)$
Bias of technical change	$\frac{\dot{s}_i}{s_i} = \frac{1}{t}\left(\alpha_{it}/s_i\right)$

apparent from Table 2 is that the share equations s_i are linear in the parameters of the function. This fact means that the parameters of a system of input demand equations represented by $(n - 1)$ share equations (one equation is redundant since $\Sigma s_i = 1$) can be estimated using standard multivariate econometric estimation techniques. The addition of the cost function to the system does not complicate the estimation technique. This addition is necessary if all the economic effects listed in Table 2 are to be estimated.

Concluding Remarks

This entry has presented the theoretical rationale for the cost function approach to production theory – that, conditional on cost minimizing behaviour, production characteristics can be represented equivalently by production functions or cost functions. Since the cost function approach is often a simpler, more direct approach to production theory, this equivalence has led to an explosion of applications of the cost function methodology in the last fifteen years. This entry has only scratched the surface of the range of topics where the cost function approach has been exploited. Important applications include aggregation theory, the economic theory of index numbers, general equilibrium theory – especially as applied in the international trade literature, public finance topics such as optimal taxation, and industrial organization topics such as the existence of natural monopoly. More information concerning the above topics and detailed discussions of duality theory can be found in Fuss and McFadden (1978) and the survey articles Diewert (1974), (Diewert 1982), Nadiri (1982) and Jorgenson (1986).

See Also

- ▶ Cost Functions
- ▶ Duality
- ▶ Joint Production
- ▶ Supply Functions

Bibliography

Baumol, W.J. 1977. *Economic theory and operations analyses*. 4th ed. Englewood Cliffs: Prentice-Hall.

Berndt, E.R., and M.A. Fuss, eds. 1986. The Econometrics of Temporary Equilibrium, a special issue of the *Journal of Econometrics*, November.

Berndt, E.R., and M. Khaled. 1979. Parametric productivity measurement and choice among flexible functional forms. *Journal of Political Economy* 87: 1220–1245.

Berndt, E.R., and D.O. Wood. 1975. Technology, prices and the derived demand for energy. *Review of Economics and Statistics* 57(3): 259–268.

Brown, R.S., D.W. Caves, and L.R. Christensen. 1979. Modelling the structure of cost and production for multiproduct firms. *Southern Economic Journal* 46: 256–270.

Burgess, D.F. 1974. Production theory and the derived demand for imports. *Journal of International Economics* 4: 103–117.

Christensen, L.R., D.W. Jorgenson, and L.J. Lau. 1973. Transcendental logarithmic production frontiers. *Review of Economics and Statistics* 55(1): 28–45.

Christensen, L., D. Jorgenson, and L.J. Lau. 1970. Conjugate duality and the transcendental logarithmic production function. Unpublished paper presented at the Second World Congress of the Econometric Society, Cambridge, England, September.

Diewert, W.E. 1969a. Canadian labor markets: a neo-classical econometric approach. Project for the Evaluation and Optimization of Economic Growth, Institute of International Studies Technical Report No. 20, Berkeley, University of California.

Diewert, W.E. 1969b. Functional form in the theory of production and consumer demand. Unpublished PhD thesis, Berkeley, University of California.

Diewert, W.E. 1971. An application of the Shephard duality theorem, a generalized Leontief production function. *Journal of Political Economy* 79(3): 481–507.

Diewert, W.E. 1974. Applications of duality. In *Frontiers of quantitative economics*, ed. M.D. Intriligator and D.A. Kendrick, Vol. II, 106–171. Amsterdam: North-Holland.

Diewert, W.E. 1982. Duality approaches to microeconomic theory. In *Handbook of mathematical economics*, ed. K.J. Arrow and M.D. Intriligator, Vol. II, 535–599. Amsterdam: North-Holland.

Diewert, W.E., and T.J. Wales. 1987. Flexible functional forms and global curvature conditions. *Econometrica* 55(1): 43–68.

Fuss, M.A. 1970. The time structure of technology: an empirical analysis of the 'putty-clay' hypothesis. Unpublished PhD dissertation, Berkeley, University of California.

Fuss, M.A. 1977a. The demand for energy in Canadian manufacturing: An example of the estimation of production functions with many inputs. *Journal of Econometrics* 5(1): 89–116.

Fuss, M.A. 1977b. The structure of technology over time: A model for testing the 'putty-clay' hypothesis. *Econometrica* 45(8): 1797–1821.

Fuss, M., and D. McFadden, ed. 1978. *Production economies: A dual approach to theory and applications*. Amsterdam: North-Holland.

Fuss, M., D. McFadden, and Y. Mundlak. 1978. A survey of functional forms in the economic analysis of production. In *Production economies: A dual approach to theory and applications*, ed. M. Fuss and D. McFadden, 219–268. Amsterdam: North-Holland.

Hanoch, G. 1970. Generation of new production functions through duality. Harvard Institute of Economic Research Discussion Paper No. 118, Cambridge, MA, April.

Hanoch, G. 1978. Generation of new production functions through duality. In *Production economics: A dual approach to theory and applications*, ed. M. Fuss and D. McFadden. Amsterdam: North-Holland.

Hotelling, H. 1932. Edgeworth's taxation paradox and the nature of demand and supply functions. *Journal of Political Economy* 40: 577–616.

Hotelling, H. 1935. Demand functions with limited budgets. *Econometrica* 3: 66–78.

Jorgenson, D.W. 1986. Econometric methods for modelling producer behaviour. In *Handbook of econometrics*, ed. Z. Griliches and M.D. Intriligator, Vol. 3, 1841–1915. Amsterdam: North-Holland.

Lau, L.J. 1974. Applications of duality theory: comments. In *Frontiers of quantitative economics*, ed. M.D. Intriligator and D.A. Kendrick, Vol. II, 176–199. Amsterdam: North-Holland.

Lau, L.J. 1976. A characterization of the normalized restricted profit function. *Journal of Economic Theory* 12: 131–163.

McFadden, D. 1966. Cost, revenue and profit functions: A cursory review. Working Paper No. 86, IBER, University of California at Berkeley, March.

McFadden, D. 1978. Cost, revenue and profit functions. In *Production economies: A dual approach to theory and applications*, ed. M. Fuss and D. McFadden. Amsterdam: North-Holland.

McKenzie, L. 1957. Demand theory without a utility index. *Review of Economic Studies* 24(3): 185–189.

Nadiri, M.I. 1982. Producers theory. In *Handbook of mathematical economics*, ed. K.J. Arrow and M.D. Intriligator, Vol. II, 431–490. Amsterdam: North-Holland.

Nerlove, M. 1963. Returns to scale in electricity supply. In *Measurement in economics: Studies in mathematical economics and econometrics in memory of Yehuda Grunfeld*, ed. C.F. Christ et al. Stanford: Stanford University Press.

Parks, R.W. 1971. Price responsiveness of factor utilization in Swedish manufacturing, 1870–1950. *Review of Economics and Statistics* 53(2): 129–139.

Roy, R. 1942. *De l'utilité*. Paris: Hermann.

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

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.

Varian, H. 1984. *Microeconomic analysis*. 2nd ed. New York: Norton.

Production as Indirect Exchange

Trout Rader

Relative to classical doctrines, neoclassical economics was distinguished early on by the addition of utility to the arsenal of theoretical tools, especially with respect to mathematical methods (e.g. Jevons (1871); Marshall (1890) and Walras (1874–7) as a supplement to such as Adam Smith (1776), Ricardo (1817) and Cournot (1838). In his *Éléments*, Walras conjectured that grouping or *tâtonnement* led to general equilibrium where supply equalled demand for each, even in a manycommodity, many-dimensioned world. Also utility was shown to be equivalent to abstract properties of preferences on bundles of consumer goods (Hicks and Allen 1934). For example, not only was choice by rational agents expostulated but, consistent with modern psychoanalysis or its milder variants, many non-rationalities were permitted Sonnenschein 1971 and Shafer 1974). Thereby application of economics was greatly expanded to cover non-transitive preferences.) However, the story did not end there.

Induced Preferences on Trade

With my thesis on Edgeworth's *Mathematical Psychics* (1881) I began a long study of the foundations of trade. Along these lines the tool from economic theory of interest is that which, on the basis of the underlying preferences and production technology, imputes consumer wellbeing to trades with properties in common usage by theorists (Rader 1963; 1964). The basic problem then is to evaluate feasible *trade* vectors, $T = \{t: t \in W + Y\}$, where Y is the *general* production possibility set (positive coordinates for outputs, negative for inputs) and W the initial endowments. The induced preference ordering is $s \succ_i$ t, i.e. agent i prefers proposed exchange s to trade t, if given s, the optimal attainable final consumption is preferred to all attainable consumptions given t. In turn for this to be sensible we should know that the optimum exists for all s (and t as well), which is known if the relevant set of trades is compact and either

(1) $\succ_i$ is transitive and irreflexive and the complement of $P_i(S) = \{t: t \succ_i S\}$ is open or, unexpectedly,
(2) Y is convex, compact, $\succ_i$ is irreflexive and complete (i.e. total) and $P_i(x)$ is convex, $P_i(s)$ is open as is its complement (Sonnenschein 1971).

Condition (1) follows for a consumer who evaluates final consumption by an upper semicontinuous, quasi-concave utility function. For instance, not only condition (1) but a relevant part of (2) is implied by the strong property of a concave utility function. However (2) applies much more widely, even to whole economies with social preferences where the properties except for transitivity are sometimes canonical (as will be seen), or perhaps more bizarrely even to political preferences based on such as majority rule which is well known to be intransitive (cf. Kramer 1972). By this construction there is a transparent derivation of induced preferences, with or without transitivity depending on which case (1) or (2) applies (Rader 1978a). In effect, *for every economy with exchange, there are preferences on trades whereupon given convexity and such, we can theorize for the whole as though there were only pure exchange*. The construction is such that basic ordering, continuity, and convexity properties can be derived and results known for exchange economies go over even if production is applied. Consequently many results of welfare economies of pure exchange are imputed to production – exchange economies by the foregoing *Principle of Equivalence*, e.g. implied is the old Fundamental Theorem of Welfare Economics that shows the equivalence of equilibrium with Pareto optimality (Gale and Mas-Colell 1977 or the newer result, Edgeworth's conjecture stated in his *Mathematical Psychics* to the effect that equilibrium is exactly that state for a large economy which does no worse for a subgroup than the group's best in autarky (Scarf 1962; Aumann 1964; Debreu and Scarf 1963). As such the Equivalence is a unification of otherwise apparently disparate materials.

Application to Hedonic Theory

This and the next three sections present applications. First, induced preferences can be applied in a Lancaster (1966) set-up where goods are demanded not for their own sake but for their underlying characteristics. Then preferences on goods are not inherent but in turn, induced. Therefore there is the sequence,

$$\text{production} + \text{initial wealth} \rightarrow \text{characteristics} \rightarrow \text{consumption},$$

and preferences are (potentially) defined in reverse at each stage. Furthermore at each stage, properties are inherited from those preceding. Mathematically, except for elementary and straightforward (if at times unexpected) syllogisms, the main relevant tools are those of the theory of linear topological spaces and more commonly finite dimensional vector spaces. This involves among other things the generalized matrix inverse used in statistics (Graybill 1969), and the formal equality of algebraic internality and topological interiority (Kelley and Namioka 1963).

Pairwise Optimality

In the case of bilateral trade, the achievement of optimality among pairs, *pairwise optimality*, often implies Pareto or consumer efficiency. By our analysis, this will apply to consumerproducers as well as those with fixed endowments. The basic requirement is interconnectedness of potential trade gains and quasi-concavity of induced utility, ore more generally convexity of induced preferences (see Rader 1968; also Polterovich 1970; and Goldman and Starr 1982, or such as ensured by the existence of a 'money good' always of use to all consumers (cf. Feldman 1973). (A unified treatment using Helly's theorem for sets was offered by Madden in 1975.) In cases of non-convexity there will still be mutual tangency of indifference hypersurfaces. By smoothness of underlying preferences, the induced preferences are smooth and by definition, there the tangent hyperplane is uniquely determined, a property inherited from the earlier stages (Rader 1976). The normal to this hyperplane defines prices and for optimality they must be equal for different consumers (since otherwise there would be exploited gains to supplementary trade). However, there is nothing to ensure that the consequent value of final consumption does or even can equal the value of initial wealth. Hence optimal trades may entail Pareto superior gifts, a subject of current research interest especially with regard to inter-country transfers, such as reparations or foreign aid (J.M. Keynes 1919; Leontief 1936; Samuelson 1952, 1954).

World Trade and Transfers

A direct application is to countries in international commerce, following Meade (1955) or Chipman (1979), each with community preferences of his own. Under a regime of free trade not only are prices equalized of final goods between countries but often so are per-unit factor costs, even for those that are immobile, so called factor price equalization (Rader 1977, 1978b; (I now prefer the phrase factor cost equalization). Therefore results for distribution of factor shares are implied as well.

The theory described above is at a relatively general level. However, there are cases of special interest. For example, if all consumers have gross substitutable demand so that an increase in one price always increases excess demand for the other goods then, even though not social utility maximizing, the economy's excess demand locally satisfies the weak law of revealed preference,

$$\delta P \Delta e < 0$$

where $e =$ excess demand and $p =$ generalized price (the vector p includes not only prices of manufactured goods but also unit costs of productive factors.)

Suppose e is chosen at p, $\bar{e}$ at $\overline{p}$. Then the weak axiom of revealed preference (WA) is that e affordable at $\overline{p}$ implies $\bar{e}$ is too expensive at p ($p\bar{e} > 0 = pe$).

Specifically, suppose the economy is partitioned into two sectors, one for consumers with gross substitutable demand and the other consisting of profit maximizing firms, then *near any equilibrium* (WA) *holds*, as stated in Rader (1972). Furthermore in this instance (WA) is easily verified for all other p near equilibrium. Also the weak law (WA) follows if consumer-traders of a given taste class are distributed with respect to expenditures over a line interval beginning at zero *expense* and ending at zero *frequency*, the frequency always non-increasing (Hildenbrand 1983). In still another case, the economy behaves as if it were one transitive consumer. It is required that all traders are homothetic and that expenditures are in constant proportions, one to the other, and the strong law (SA) is said to hold (Chipman 1974).

In these various instances, the whole economy acts like a single representative consumer with various rationality properties. The Marshallian idealization where marginal utility of income or money is assumed constant so that the whole acts like a single consumer is more than metaphor that permits aggregation of consumer demand. The 'offer' curve is indeed the result of social optimization and also choice is still the intersection of the community feasibility set with a social indifference hypersurface. It is only that disjoint surfaces may intersect, reflecting the absence of transitivity.

Transaction Costs

One particular technology which was mentioned in Debreu (1959, ch. 7) has, evidently independently, received much attention and continues to have prospects, namely that of the technology of transferring a good from one location to another or more generally, transaction costs (cf. Coase 1960). Since the conditions sufficient for induced preferences to be upper-semi continuous are very weak, the set of Pareto optima is very likely non-empty. However, if the technical possibilities do not form a convex set, whence induced preferences on trades may not be convex even when those for direct consumption are, then the welfare equivalence between optimality and equilibrium may not be, whence non-competition may be appropriate.

See Also

- ▶ Arrow–Debreu Model of General Equilibrium
- ▶ General Equilibrium

Bibliography

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

Aumann, R.J. 1966. Existence of competitive equilibria in markets with a continuum of traders. *Econometrica* 34: 1–17.

Birkhoff, G., and G.C. Rota. 1962. *Ordinary differential equations*. Boston: Ginn & Co.

Chipman, J. 1974. Homothetic preferences and aggregation. *Journal of Economic Theory* 8(1): 26–38.

Chipman, J. 1979. The theory and application of trade utility functions. In *General equilibrium, growth, and trade*, ed. J.R. Green and J.A. Scheinkman. New York: Academic Press.

Coase, R.H. 1960. The problem of social cost. *Journal of Law and Economics* 3: 1–44.

Coddington, E.A., and N. Levinson. 1955. *Theory of ordinary differential equations*. New York: McGraw-Hill.

Cournot, A. 1838. *Recherches sur les principes mathématiques de la théorie des richesses*. Trans. New York: Macmillan, 1983.

Debreu, G. 1959. *Theory of value*. New York: Wiley.

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

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

Feldman, A.M. 1973. Bilateral trading processes, pairwise optimality and Pareto optimality. *Review of Economic Studies* 40: 463–479.

Gale, D., and A. Mas-Colell. 1975. An equilibrium existence theorem for a general model without ordered preferences. *Journal of Mathematical Economics* 2: 9–15.

Gale, D., and A. Mas-Colell. 1976–7. On the role of complete, transitive preferences in equilibrium theory. In *Equilibrium and disequilibrium in economics*, ed. G. Schwodiauer. Dordrecht: Reidel.

Gantmacher, F.R. 1959. *The theory of matrices*, vol. 1. New York: Chelsea.

Girsanov, I.V. 1972. *Lectures on mathematical extremum problems*. New York: Springer.

Goldman, S., and R. Starr. 1982. Pairwise, t-wise, and Pareto optimalities. *Econometrica* 50: 593–606.

Graybill, F. 1969. *Introduction to matrices*. Belmont: Wadsworth.

Hicks, J.R., and R.G.D. Allen. 1934. A reconsideration of the theory of value. *Economica* NS 1: 52–76.

Hildenbrand, W. 1983. On the law of demand. *Econometrica* 51: 997–1019.

Jevons, W.S. 1871. *The theory of political economy.* London: Macmillan.

Kelley, J., and I. Namioka. 1963. *Linear topological spaces*. New York: Van Nostrand.

Keynes, J.M. 1919. *The economic consequences of the peace*. London: Macmillan.

Kihlstrom, R., A. Mas-Colell, and H. Sonnenschein. 1976. The demand theory of the weak axiom of revealed preference. *Econometrica* 44: 971–978.

Kramer, G.H. 1972. Sophisticated voting over multidimensional choice spaces. *Journal of Mathematical Sociology* 2: 165–180.

Lancaster, K. 1966. A new approach to consumer theory. *Journal of Political Economy* 74: 132–157.

Leontief, W. 1936. Note on the pure theory of transfer. In *Explorations in economics*. New York: McGraw-Hill.

Leontief, W. 1941. *The structure of American economy, 1919–1939*, 1951. London/New York: Oxford University Press.

Madden, P. 1975. Efficient sequences of non-monetary exchange. *Review of Economic Studies* 44: 581–596.

Marshall, A. 1890. *Principles of economics*. London: Macmillan.

Mas-Colell, A. 1974. An equilibrium existence theorem without complete or transitive preferences. *Journal of Mathematical Economics* 1: 237–246.

McKenzie, L. 1960. Matrices with dominant diagonals and economic theory. In *Mathematical methods in the social sciences*, ed. K. Arrow, S. Karlin, and P. Suppes. Stanford: Stanford University Press.

Meade, J.E. 1955. *Trade and welfare*. Oxford: Oxford University Press.

Polterovich, M. 1970. A model of resource distribution. *Matekon* 7(3): 245–262.

Rader, T. 1963. *Edgeworth exchange and general economic equilibrium*. Unpublished dissertation, Yale University. Synopsis in Yale Economic Essays 4, 1964, 133–80.

Rader, T. 1968. Pairwise optimality and non-competitive behavior. In *Papers in quantitative economics*, vol. I, ed. J. Quirk and A.M. Zarley. Lawrence: University of Kansas Press.

Rader, T. 1970. Resource allocation with increasing returns to scale. *American Economic Review* 60: 814–825.

Rader, T. 1972. General equilibrium theory with complementary factors. *Journal of Economic Theory* 4: 372–380.

Rader, T. 1976. Pairwise optimality, multilateral optimality and efficiency, with and without externalities. In *Theory and measurement of economic externalities*, ed. S.Y. Lin. New York: Academic Press.

Rader, T. 1977–8. Many good multiplier analysis: Classical, neoclassical and Keynesian. In *Equilibrium and Disequilibrium in Economics*, ed. G. Schwodiauer. Dordrecht: Reidel Press.

Rader, T. 1978a. Induced preferences on trades when preferences may be intransitive and incomplete. *Econometrica* 46: 137–146.

Rader, T. 1978b. On factor price equalization. *Journal of Mathematical Economics* 5: 71–82.

Rader, T. 1979. Factor price equalization with more industries than factors. In *General equilibrium growth and trade*, ed. J. Green and J.A. Scheinkman. New York: Academic Press.

Ricardo, D. 1817. *The principles of political economy and taxation*. London: Murray.

Samuelson, P.A. 1952–4. The transfer problem and transport costs. I: The terms of trade when impediments are absent; II: Analysis of effects of trade impediments. *Economic Journal* 57: 278–304; 59: 264–90.

Scarf, H. 1962. An analysis of markets with a large number of participants. In *Recent advances in game theory*, ed. H. Scarf. Princeton: Princeton University Press.

Shafer, W. 1974. The nontransitive consumer. *Econometrica* 42(5): 913–919.

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

Sonnenschein, H. 1971. Demand theory without transitive preferences. In *Preferences, utility and demand*, ed. J. Chipman, M. Hurwicz, M. Richter, and H. Sonnenschein. New York: Harcourt, Brace.

Walras, L. 1874–7. *Eléments d'économie politique pure*. Lausanne: L. Corbaz. Definitive edn 1926, trans. W. Jaffé as Elements of pure economics. New York: Orion, 1954.

Production Functions

Dale W. Jorgenson

Abstract

Traditionally, the production function was assumed to be additive and homogeneous. The constant elasticity of substitution (CES) production function adds flexibility by treating the elasticity of substitution as an unknown parameter, but retains the assumptions of additivity and homogeneity and imposes very stringent limitations on patterns of substitution. The dual formulation of production theory characterizes the production function by means of a dual representation such as a price or cost function, and generates explicit demand and supply functions as derivatives of the price or cost function.

Keywords

Bias of technical change; CES production function; Cobb–Douglas functions; Cost flexibility; Cost function; Demand function; Economies of scale; Elasticity of substitution; Hicks, J.; Implicit function theorem; Input–output analysis; Jorgensen, D. W.; Price function; Production functions; Simultaneous equations models; Supply function; Technical change

JEL Classifications
D2

Introduction

The economic theory of production – as presented in such classic treatises as Hicks's *Value and Capital* (1946) and Samuelson's *Foundations of Economic Analysis* (1983) – is based on the maximization of profit, subject to a production function. The objective of this theory is to characterize demand and supply functions, using only the restrictions on producer behaviour that arise from optimization. The principal analytical tool employed for this purpose is the implicit function theorem.

The traditional approach to economic modelling of producer behaviour begins with the assumption that the production function is additive and homogeneous.

Under these restrictions demand and supply functions can be derived explicitly from the production function and the necessary conditions for producer equilibrium. However, this approach has the disadvantage of imposing constraints on patterns of producer behaviour – thereby frustrating the objective of determining these patterns empirically.

The traditional approach was originated by Cobb and Douglas (1928) and was employed in empirical research by Douglas (1948, 1967, 1976) and his associates for almost two decades. The limitations of this approach were made strikingly apparent by Arrow et al. (1961, henceforward ACMS), who pointed out that the Cobb–Douglas production function imposes a priori restrictions on patterns of substitution among inputs. In particular, elasticities of substitution among all inputs must be equal to unity.

The constant elasticity of substitution (CES) production function introduced by ACMS adds flexibility to the traditional approach by treating the elasticity of substitution as an unknown parameter. However, the CES production function retains the assumptions of additivity and homogeneity and imposes very stringent limitations on patterns of substitution. McFadden (1963) and Uzawa (1962) have shown, essentially, that elasticities of substitution among all inputs must be the same.

The dual formulation of production theory has made it possible to overcome the limitations of the traditional approach to econometric modelling. This formulation was introduced by Hotelling (1932) and later revived and extended by Samuelson (1953, 1960) and Shephard (1953, 1970). The key features of the dual formulation are, first, to characterize the production function by means of a dual representation such as a price or cost function, and second, to generate explicit demand and supply functions as derivatives of the price or cost function.

Patterns of producer behaviour can be described most usefully in terms of the behaviour of the derivatives of demand and supply functions. For example, measures of substitution can be specified in terms of the response of demand patterns to changes in input prices. Similarly, measures of technical change can be specified in terms of the response of these patterns to changes in technology. The classic formulation of production theory at this level of specificity can be found in Hicks's *Theory of Wages* (1963).

Hicks (1963) introduced the elasticity of substitution as a measure of substitutability. The elasticity of substitution is the proportional change in the ratio of two inputs with respect to a proportional change in their relative price. Similarly, Hicks introduced the bias of technical change as a measure of the impact of changes in technology on patterns of demand for inputs. The bias of technical change is the response of the share of an input in the value of output to a change in the level of technology.

By treating measures of substitution and technical change as fixed parameters the system of demand and supply functions can be generated by integration. Provided that the resulting functions are themselves integrable, the underlying price or cost function can be obtained by a second integration. As we have already pointed out, Hicks's elasticity of substitution is unsatisfactory for this purpose, since it leads to arbitrary restrictions on patterns of producer behaviour.

The introduction of a new measure of substitution, the share elasticity, by Christensen et al. (1971, 1973) and Samuelson (1973) has made it possible to overcome the limitations of parametric forms based on constant elasticities of substitution. Share elasticities, like biases of technical change, can be defined in terms of shares of inputs in the value of output. The share elasticity of a given input is the response of the share of that input to a proportional change in the price of an input.

Models of Producer Behaviour

The purpose of this section is to present the simplest form of the economic theory of production. We base this theory on a production function with constant returns to scale. Producer equilibrium implies the existence of a price function, giving the price of output as a function of the prices of inputs and the level of technology. The price function is dual to the production function and provides an alternative and equivalent description of technology.

An econometric model of producer behaviour takes the form of a system of simultaneous equations, determining the distributive shares of the inputs and the rate of technical change as functions of the input prices and the level of technology. Measures of substitution and technical change give the responses of the distributive shares and the rate of technical change to changes in prices and technology. To generate an econometric model of producer behaviour we treat these measures as unknown parameters to be estimated.

In order to present the theory of production we first require some notation. We denote the quantity of output by y and the quantities of J inputs by $x_j(j = 1, 2, \ldots, J)$. Similarly, we denote the price of output by q and the prices of the J inputs by $p_j(j = 1, 2, \ldots, J)$. We find it convenient to employ vector notation for the input quantities and prices:

$x = (x_1, x_2, \ldots, x_J)$ – vector of input quantities.

$p = (p_1, p_2, \ldots, p_J)$ – vector of input prices.

We assume that the technology can be represented by a *production* function, say *F*, where:

$$y = F(x, t), \tag{1}$$

and t is an index of the level of technology. In the analysis of time series data for a single producing unit the level of technology can be represented by time. In the analysis of cross-section data for different producing units the level of technology can be represented by one-zero dummy variables corresponding to the different units. We can define the *shares* of inputs in the value of output by:

$$v_j = \frac{p_j x_j}{qy}, (j = 1, 2, \ldots, J).$$

Under competitive markets for output and all inputs the necessary conditions for producer equilibrium are given by equalities between the share of each input in the value of output and the elasticity of output with respect to that input:

$$v = \frac{\partial \ln y(x, t)}{\partial \ln x}, \tag{2}$$

where:

$v = (v_1, v_2, \ldots, v_J)$ – vector of value shares.

$\ln x = (\ln x_1, \ln x_2, \ldots, \ln x_J)$ – vector of logarithms of input quantities.

Under constant returns to scale the elasticities and the value shares for all inputs sum to unity:

$$i'v = i' \frac{\partial \ln y}{\partial \ln x} = 1,$$

where i is a vector of ones. The value of output is equal to the sum of the values of the inputs.

Finally, we can define the *rate of technical change*, say v_t, as the rate of growth of the quantity of output holding all inputs constant:

$$v_t = \frac{\partial \ln y(x, t)}{\partial t}. \tag{3}$$

It is important to note that this definition does not impose any restriction on patterns of substitution among inputs.

Given the identity between the value of output and the value of all inputs and given equalities between the value share of each input and the elasticity of output with respect to that input, we can express the price of output as a function, say *Q,* of the prices of all inputs and the level of technology:

$$q = Q(p, t). \tag{4}$$

We refer to this as the *price* function for the producing unit.

The price function Q is dual to the production function F and provides an alternative and equivalent description of the technology of the producing unit. We can formalize this description in terms of the following properties of the price function:

(1) *Positivity.* The price function is positive for positive input prices.
(2) *Homogeneity.* The price function is homogeneous of degree one in the input prices.
(3) *Monotonicity.* The price function is increasing in the input prices.
(4) *Concavity.* The price function is concave in the input prices.

Given differentiability of the price function, we can express the value shares of all inputs as elasticities of the price function with respect to the input prices:

$$v = \frac{\partial \ln q(p, t)}{\partial \ln p}, \tag{5}$$

where:

ln p = (lnp$_1$, lnp$_2$, lnp$_J$) – vector of logarithms of input prices.

Since the price function is increasing in the input prices the value shares must be non-negative.

We can express the negative of the rate of technical change as the rate of growth of the price of output, holding the prices of all inputs constant:

$$-v_t = \frac{\partial \ln q(p, t)}{\partial t}. \tag{6}$$

Since the price function Q is homogeneous of degree one in the input prices, the value shares and the rate of technical change are homogeneous of degree zero and the value shares sum to unity.

We have represented the value shares of all inputs and the rate of technical change as functions of the input prices and the level of technology. We can introduce measures of substitution and technical change to characterize these functions in detail. For this purpose we differentiate the logarithm of the price function twice with respect to the logarithms of input prices to obtain measures of substitution:

$$U_{pp} = \frac{\partial^2 \ln q(p, t)}{\partial \ln p^2} = \frac{\partial v(p, t)}{\partial \ln p}. \tag{7}$$

We refer to the measures of substitution (7) as *share elasticities,* since they give the response of the value shares of all inputs to proportional changes in the input prices. If a share elasticity is positive, the corresponding value share increases with the input price. If a share elasticity is negative, the value share decreases with the input price. Finally, if a share elasticity is zero, the value share is independent of the price.

Second, we can differentiate the logarithm of the price function twice with respect to the logarithms of input prices and the level of technology to obtain measures of technical change:

$$u_{pt} = \frac{\partial^2 \ln q(p, t)}{\partial \ln p \partial t} = \frac{\partial v}{\partial t} = -\frac{\partial v_t(p, t)}{\partial \ln p}. \tag{8}$$

We refer to these measures as *biases of technical change.* If a bias of technical change is positive, the corresponding value share increases with a change in the level of technology and we say that technical change is *input-using.* If a bias of technical change is negative, the value share decreases with a change in technology and technical change is *input-saving.* Finally, if a bias is zero, the value share is independent of technology; in this case we say that technical change is *neutral* (in the sense of Hicks).

Alternatively, the vector of biases of technical change u_{pt} can be employed to derive the implications of changes in input prices for the rate of technical change. If a bias of technical change is positive, the rate of technical change decreases with the input price. If a bias is negative, the rate of technical change increases with the input price. Finally, if a bias is zero so that technical change is neutral, the rate of technical change is independent of the price.

To complete the description of technical change we can differentiate the logarithm of the price function twice with respect to the level of technology:

$$u_{tt} = \frac{\partial^2 \ln q(p,t)}{\partial t^2} = -\frac{\partial v_t(p,t)}{\partial t} \tag{9}$$

We refer to this measure as the *deceleration* of technical change, since it is the negative of rate of change of the rate of technical change. If the deceleration is positive, negative, or zero, the rate of technical change is decreasing, increasing, or independent of the level of technology.

The matrix of second-order logarithmic derivatives of the logarithm of the price function Q must be symmetric. This matrix includes the matrix of share elasticities U_{pp}, the vector of biases of technical changes u_{pt}, and the deceleration of technical change u_{tt}. Concavity of the price function in the input prices implies that the matrix of second-order derivatives, say H, is nonpositive definite, so that the matrix $U_{pp} + vv' - V$ is nonpositive definite, where:

$$\frac{1}{q} N \cdot H \cdot N = U_{pp} + vv' - V;$$

the price of output q is positive and the matrices N and V are diagonal:

$$N = \begin{bmatrix} p_1 & 0 & \ldots & 0 \\ 0 & p_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & p_J \end{bmatrix}, \quad V = \begin{bmatrix} \upsilon_1 & 0 & \ldots & 0 \\ 0 & \upsilon_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \upsilon_J \end{bmatrix}$$

We can define substitution and complementarity of inputs in terms of the matrix of share elasticities U_{pp} and the vector of value shares v. We say that two inputs are *substitutes* if the corresponding element of the matrix $U_{pp} + vv' - V$ is negative. Similarly, we say that two inputs are *complements* if the corresponding element of this matrix is positive. If the element of this matrix corresponding to the two inputs is zero, we say that the inputs are *independent.* The definition of substitution and complementarity is symmetric in the two inputs, reflecting the symmetry of the matrix $U_{pp} + vv' - V$. If there are only two inputs, nonpositive definiteness of this matrix implies that the inputs cannot be complements.

To generate an econometric model of producer behaviour a natural approach is to treat the measures of substitution and technical change as unknown parameters to be estimated. For this purpose we introduce the parameters:

$$B_{pp} = U_{pp}, \beta_{pt} = u_{pt}, \beta_{tt} = u_{tt}, \tag{10}$$

where B_{pp} is a matrix of constant share elasticities, β_{pt} is a vector of constant biases of technical changes, and β_{tt} is a constant deceleration of technical change.

We can regard the matrix of share elasticities, the vector of biases of technical change, and the deceleration of technical change as a system of second-order partial differential equations. We can integrate this system to obtain a system of first-order partial differential equations:

$$\begin{aligned} v &= \alpha_p + B_{pp} \ln p + \beta_{pt} \cdot t, \\ -v_t &= \alpha_t + \beta'_{pt} \ln p + \beta_{tt} \cdot t, \end{aligned} \tag{11}$$

where the parameters –α_p, α– are constants of integration.

To provide an interpretation of the parameters – α_p, α_t – we first normalize the input prices. We can set the prices equal to unity where the level of technology t is equal to zero. This represents a choice of origin for measuring the level of technology and a choice of scale for measuring the quantities and prices of inputs. The vector of parameters α_p is the vector of value shares and the parameter α_t is the negative of the rate of technical change where the level of technology t is zero.

Similarly, we can integrate the system of first-order partial differential equations (11) to obtain the price function:

$$\ln p = \alpha_0 + \alpha_p' \ln p + \alpha_t \cdot t + \frac{1}{2} \ln p' B_{pp} \ln p \\ + \ln p' \beta_{pt} \cdot t + \frac{1}{2} \beta_{tt} \cdot t^2, \qquad (12)$$

where the parameter α_0 is a constant of integration. Normalizing the price of output so that it is equal to unity where t is zero, we can set this parameter equal to zero. This represents a choice of scale for measuring the quantity and price of output.

For the price function (12) the price of output is a transcendental or, more specifically, an exponential function of the logarithms of the input prices. We refer to this form as the *transcendental logarithmic* price function or, more simply, the translog price function, indicating the role of the variables. We can also characterize this price function as the *constant share elasticity* or CSE price function, indicating the role of the fixed parameters. In this representation the scalars – α_t, β_{tt} – the vectors – α_p, β_{pt} – and the matrix B_{pp} are constant parameters that reflect the underlying technology. Differences in levels of technology among time periods for a given producing unit or among producing units at a given point of time are represented by differences in the level of technology t.

Economies of Scale

In section "Models of Producer Behaviour" we have considered producer behaviour under constant returns to scale. In this section we consider producer behaviour under increasing returns to scale. Under increasing returns and competitive markets for output and all inputs, producer equilibrium is not defined by profit maximization, since no maximum of profit exists. However, in regulated industries the price of output is set by regulatory authority. Given demand for output as a function of the regulated price, the level of output is exogenous to the producing unit.

With output fixed from the point of view of the producer, necessary conditions for equilibrium can be derived from cost minimization. Where total cost is defined as the sum of expenditures on all inputs, the minimum value of cost can be expressed as a function of the level of output and the prices of all inputs. We refer to this function as the cost function. We have described the theory of production under constant returns to scale in terms of properties of the price function (3); similarly, we can describe the theory under increasing returns in terms of properties of the cost function.

Utilizing the notation of section "Models of Producer Behaviour", we can define total cost, say c, as the sum of expenditures on all inputs:

$$c = \sum_{j=1}^{J} p_j x_j.$$

We next define the shares of inputs in total cost by:

$$v_j = \frac{p_j x_j}{c}, (j = 1, 2, \ldots, J).$$

With output fixed from the point of view of the producing unit and competitive markets for all inputs, the necessary conditions for producer equilibrium are given by equalities between the shares of each input in total cost and the ratio of the elasticity of output with respect to that input and the sum of all such elasticities:

$$v = \frac{\dfrac{\partial \ln y}{\partial \ln x}}{i' \dfrac{\partial \ln y}{\partial \ln x}}, \qquad (13)$$

where i is a vector of ones and:

$$v = (v_1, v_2, \ldots, v_J) - \text{vector of cost shares.}$$

Given the definition of total cost and the necessary conditions for producer equilibrium, we can express total cost, say c, as a function of the prices of all inputs and the level of output:

$$c = C(p, y). \qquad (14)$$

We refer to this as the *cost* function. The cost function C is dual to the production function F and provides an alternative and equivalent description of the technology of the producing unit.

We can formalize the theory of production in terms of the following properties of the cost function:

(1) *Positivity.* The cost function is positive for positive input prices and a positive level of output.
(2) *Homogeneity.* The cost function is homogeneous of degree one in the input prices.
(3) *Monotonicity.* The cost function is increasing in the input prices and in the level of output.
(4) *Concavity.* The cost function is concave in the input prices.

Given differentiability of the cost function, we can express the cost shares of all inputs as elasticities of the cost function with respect to the input prices:

$$v = \frac{\partial \ln c(p, y)}{\partial \ln p}. \tag{15}$$

Since the cost function is increasing in the input prices, the cost shares must be nonnegative.

We can define an index of returns to scale as the elasticity of the cost function with respect to the level of output:

$$v_y = \frac{\partial \ln c}{\partial \ln y}(p, y). \tag{16}$$

Following Frisch (1965), we can refer to this elasticity as the *cost flexibility.* The cost function is increasing in the level of output, so that the cost flexibility is positive. Since the cost function C is homogeneous of degree one in the input prices, the cost shares and the cost flexibility are homogeneous of degree zero and the cost shares sum to unity.

The cost flexibility v_y is the reciprocal of the *degree of returns to scale*, defined as the elasticity of output with respect to a proportional increase in all inputs:

$$v_y = \frac{1}{i' \dfrac{\partial \ln y}{\partial \ln x}}. \tag{17}$$

If output increases more than in proportion to the increase in inputs, cost increases less than in proportion to the increase in output.

We have represented the cost shares of all inputs and the cost flexibility as functions of the input prices and the level of output. We can characterize these functions in terms of measures of substitution and economies of scale. We obtain *share elasticities* by differentiating the logarithm of the cost function twice with respect to the logarithms of input prices:

$$U_{pp} = \frac{\partial^2 \ln c(p, y)}{\partial \ln p^2} = \frac{\partial v(p, y)}{\partial \ln p}. \tag{18}$$

These measures of substitution give the response of the cost shares of all inputs to proportional changes in the input prices.

Second, we can differentiate the logarithm of the cost function twice with respect to the logarithms of the input prices and the level of output to obtain measures of economies of scale:

$$\begin{aligned} u_{py} &= \frac{\partial^2 \ln c(p, y)}{\partial \ln p \partial \ln y} = \frac{\partial v(p, v)}{\partial \ln y} \\ &= -\frac{\partial v_y(p, y)}{\partial \ln p}. \end{aligned} \tag{19}$$

We refer to these measures as *biases of scale.* The vector of biases of scale u_{py} can be employed to derive the implications of economies of scale for the relative distribution of total cost among inputs. Alternatively, this vector can be employed to derive the implications of changes in input prices for the cost flexibility. To complete the description of economies of scale we can differentiate the logarithm of the cost function twice with respect to the level of output:

$$u_{yy} = \frac{\partial^2 \ln c(p, y)}{\partial \ln y^2} = \frac{\partial v_y(p, y)}{\partial \ln y}. \tag{20}$$

The matrix of second-order logarithmic derivatives of the logarithms of the cost function C must be symmetric. This matrix includes the matrix of share elasticities U_{pp}, the vector of biases of scale u_{py}, and the derivative of the cost flexibility with respect to the logarithm of output u_{yy}. Concavity of the cost function in the input prices implies that the matrix of second-order derivatives, say H, is nonpositive definite, so that the matrix $U_{pp} + vv' - V$ is nonpositive definite, where:

$$\frac{1}{c} N \cdot H \cdot N = U_{pp} + vv' - V.$$

Total cost c is positive and the diagonal matrices N and V are defined in terms of the input prices p and the cost shares v, as in section "Models of Producer Behaviour".

We say that the cost function C is *homothetic* if and only if the cost function is separable in the prices of all J inputs $\{p_1, p_2, \ldots, p_J\}$, so that:

$$c = C[P(p_1, p_2, \ldots, p_J), y], \tag{21}$$

where the function P is homogeneous of degree one and independent of the level of output y. The cost function is homothetic if and only if the production function is *homothetic,* where:

$$y = F[G(x_1, x_2, \ldots, x_J)], \tag{22}$$

where the function G is homogeneous of degree one.

Since the cost function is homogeneous of degree one in the input prices, it is homogeneous of degree one in the function P, which can be interpreted as the price index for a single aggregate input; the function G is the corresponding quantity index. Furthermore, the cost function can be represented as the product of the price index of aggregate input P and a function, say H, of the level of output:

$$c = P(p_1, p_2, \ldots, p_J) \cdot H(y). \tag{23}$$

Under homotheticity, the cost flexibility v_y is independent of the input prices:

$$v_y = \frac{\partial \ln H}{\partial \ln y}(y). \tag{24}$$

If the cost flexibility is also independent of the level of output, the cost function is homogeneous in the level of output and the production function is homogeneous in the quantity index of aggregate input G. The degree of homogeneity of the production function is the degree of returns to scale and is equal to the reciprocal of the cost flexibility. Under constant returns to scale the degree of returns to scale and the cost flexibility are equal to unity.

We can generate an econometric model of cost and production by introducing the parameters:

$$B_{pp} = U_{pp}, \beta_{py} = u_{py}, \beta_{yy} = u_{yy}, \tag{25}$$

where B_{pp} is a matrix of constant share elasticities, β_{py} is a vector of constant biases of scale, and β_{yy} is a constant derivative of the cost flexibility with respect to the logarithm of output. We can treat the matrix of constant parameters as a system of second-order partial differential equations, obtaining:

$$\begin{aligned} v = \alpha_p + B_{pp} \ln p + \beta_{py} \ln y,\ v_y \\ = \alpha_y + \beta'_{py} \ln p + \beta_{yy} \ln y, \end{aligned} \tag{26}$$

where the parameters – α_p, α_y – are constants of integration.

We can integrate the system (26) to obtain the cost function:

$$\begin{aligned} \ln c = \alpha_0 + \alpha_p \ln p + \alpha_y + \frac{1}{2} \ln p' B_{pp} \ln p \\ + \ln p' \beta_{py} \ln y + \frac{1}{2} \beta_{yy} (\ln y)^2, \end{aligned} \tag{27}$$

where the parameter $-\alpha_0$ is a constant of integration. We can refer to this form as the translog cost function, indicating the role of the variables, or the constant share elasticity (CSE) cost function, indicating the role of the parameters.

Under homotheticity the cost flexibility is independent of the input prices. A necessary and sufficient condition for homotheticity is given by:

$$\beta_{py} = 0; \tag{28}$$

the vector of biases of scale is equal to zero. Under homogeneity the cost flexibility is also independent of output, so that:

$$\beta_{yy} = 0;$$

the derivative of the flexibility with respect to the logarithm of output is zero. Finally, under constant returns to scale, the cost flexibility is equal to unity; given the restrictions implied by homotheticity, constant returns requires:

$$\alpha_y = 1. \qquad (29)$$

Summary and Conclusion

The econometric modelling of producer behaviour requires parametric forms for demand and supply functions. Patterns of production can be represented in terms of unknown parameters that specify the responses of demands and supplies to changes in prices, technology and scale. New measures of substitution, technical change and economies of scale have provided greater flexibility in the empirical determination of production patterns. These innovations have arisen from the dual formulation of the theory of production.

We can conclude by suggesting possible directions for future research. The primary focus of our discussion has been on the characterization of technology for individual producing units. Application of the results typically involves models for both demand and supply of a given commodity. The ultimate objective of econometric modelling of production is to construct general equilibrium models encompassing demands and supplies for a wide range of products and factors of production, along the lines suggested by Jorgenson (1983).

Our exposition of the theory of production has emphasized models where the econometric methodology has crystallized. An important area for future research is the implementation of dynamic models of technology. These models are based on substitution possibilities among outputs and inputs at different points of time. The simplest intertemporal model of production is based on capital as a factor of production. This model is treated in a companion paper by Jorgenson (1986).

A number of promising avenues for further investigation have been suggested in the literature on the theory of production summarized in the entry on vintages.

See Also

- ▶ CES Production Function
- ▶ Cobb–Douglas Functions

Bibliography

Arrow, K.J., H.B. Chenery, B.S. Minhas, and R.M. Solow. 1961. Capital–labor substitution and economic efficiency. *Review of Economics and Statistics* 63 (3): 225–247.

Christensen, L.R., D.W. Jorgenson, and L.J. Lau. 1971. Conjugate duality and the transcendental logarithmic production function. *Econometrica* 39 (3): 255–256.

Christensen, L.R., D.W. Jorgenson, and L.J. Lau. 1973. Transcendental logarithmic production frontiers. *Review of Economics and Statistics* 55 (1): 28–45.

Cobb, C.W., and P.H. Douglas. 1928. A theory of production. *American Economic Review* 18: 139–165.

Douglas, P.H. 1948. Are there laws of production? *American Economic Review* 38: 1–41.

Douglas, P.H. 1967. Comments on the Cobb–Douglas production function. In *The theory and empirical analysis of production*, ed. M. Brown, 15–22. New York: Columbia University Press.

Douglas, P.H. 1976. The Cobb–Douglas production function once again: Its history, its testing, and some empirical values. *Journal of Political Economy* 84: 903–916.

Frisch, R. 1965. *Theory of production*. Chicago: Rand McNally. (English translation from the 9th edn of lectures published in Norwegian; the 1st edn of the lectures dates from 1926.)

Hicks, J.R. 1946. *Value and capital*. 2nd ed. Oxford: Oxford University Press.

Hicks, J.R. 1963. *The theory of wages*. 2nd ed. London: Macmillan.

Hotelling, H.S. 1932. Edgeworth's taxation paradox and the nature of demand and supply functions. *Journal of Political Economy* 40: 577–616.

Jorgenson, D.W. 1983. Modeling production for general equilibrium analysis. *Scandinavian Journal of Economics* 85 (2): 101–112.

Jorgenson, D.W. 1986. Econometric methods for modeling producer behavior. In *Handbook of econometrics*, ed. Z. Griliches and M.D. Intriligator, vol. 3. North-Holland: Amsterdam.

McFadden, D. 1963. Further results on CES production functions. *Review of Economic Studies* 30 (2): 73–83.

Samuelson, P.A. 1953. Prices of factors and goods in general equilibrium. *Review of Economic Studies* 21 (1): 1–20.

Samuelson, P.A. 1960. Structure of a minimum equilibrium system. In *Essays in economics and econometrics*, ed. R.W. Pfouts, 1–33. Chapel Hill: University of North Carolina Press.

Samuelson, P.A. 1973. Relative shares and elasticities simplified: Comment. *American Economic Review* 63: 770–771.

Samuelson, P.A. 1983. *Foundations of economic analysis*. 2nd ed. Cambridge, MA: Harvard University Press.

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. 1962. Production functions with constant elasticity of substitution. *Review of Economic Studies* 29 (4): 291–299.

Production: Neoclassical Theories

M. Ishaq Nadiri

The economic theory of production is concerned with the characterization of the input demand and output supply functions based on a theory of profit maximization subject to a production function. Two sets of issues are involved: one is the technical constraint that describes the range of production processes available to the firm, and the other is the make-up of the markets where the firm's transactions take place. There is a substantial literature on the latter, which cannot be addressed here: we adopt the admittedly unrealistic assumption of 'perfect competition' in both commodity and factor markets. Our purpose is to discuss the properties of the production technology in the context of the neoclassical theory of the multiple-product and multiple-input firm, identify the specific forms of the production function which are proposed in the literature, and discuss the duality principles as well as some of the new dynamic factor demand model analyses.

Neoclassical Theory of Production

Consider a firm that produces n products and employs m inputs; its objective is to maximize profits given as:

$$\Pi = \sum_{i=1}^{m+n} p_i y_i = \sum_{i=1}^{n} p_i y_i + \sum_{i=n+1}^{m} p_i y_i \qquad (1)$$

where $y_i(i = 1,\ldots, n)$ are the outputs and p_i $(i = \mathrm{n} + 1, \ldots, m)$ are output prices: $\mathrm{y}_{n+1} = x_i(i = 1,\ldots,m)$ are the inputs, and $p_{n+1}(i = 1,\ldots, m)$ are the input prices. Profit, Π, is maximized subject to the production function

$$f(y_1, y_2, \ldots, y_n, y_{n+1}, \ldots, y_{m+n}) = 0. \qquad (2)$$

$y_1, y_2, \ldots, y_{m+n}$ are often called *net outputs;* they have positive signs for outputs and negative signs for inputs.

Assuming that the production function $f(\cdot)$ is (1) twice differentiable, i.e.

$$\partial f/\partial y_i = f_i \text{ and } \partial^2 f/\partial y_i \partial y_j = f_{ij}, (i,j = 1, \ldots, m + n),$$

exist; (2) increasing in the *net* outputs, i.e. the derivatives, f_i are always positive; and (3) convex (subject to the condition $f(\cdot) = 0$, the function is *strictly* convex); the optimal production plan of the firm can be stated using the familiar Lagrangian function:

$$\begin{aligned} L(y_1, y_2, \ldots, y_{m+n}; \lambda) &= \Pi + \lambda f(y_i, y_2, \ldots, y_{m+n}) \\ &= \sum_{i=1}^{m+n} p_i y_i + \lambda f(y_1, y_2, \ldots, y_{m+n}), \end{aligned} \qquad (3)$$

where λ is a Lagrange multiplier associated with the constraint $f(\cdot) = 0$.

There are $m + n$ first-order conditions that can be interpreted as equality between the marginal profitability of each net output and its revenue or cost. The Lagrange multiplier is the change in profit made by the firm with respect to a change in its production plan. Manipulating these equalities, we obtain familiar expressions such as the marginal transformation among commodities and inputs, the marginal rate of technical substitution among inputs and the expansion path of inputs. It follows that the profit-maximizing output and input levels, $\bar{y}_i(i = 1,\ldots, m + n)$, and the

Lagrange multiplier, $\overline{\lambda}$ are functions of the prices $p_i(i = 1,\ldots, m + n)$. That is:

$$\overline{\lambda} = \overline{\lambda}(p_1, \ldots, p_{m+n})$$

and

$$\overline{y}_i = s^i(p_1, \ldots, p_{m+n}), \quad (i = 1, \ldots, m + n). \tag{4}$$

$\overline{\lambda}(\cdot)$ is homogeneous of degree one, while s^i $(\cdot)$ is homogeneous of degree zero. $\overline{y}_i$ are the *net supply functions*. For outputs, the equations i y are the usual supply functions; for inputs, they are the negative of the demand functions. Thus net supply functions exist provided that the marginal profitability conditions are satisfied and that the production function has the appropriate properties.

Production: Neoclassical Theories, Table 1 A partial list of economic effects related to the production function

Output level	$y = f(x, t)$
Returns to scale	$\mu = \left(\sum_{i=1}^{n} x_i f_i\right)/f$
Distributive share	$s_1 = x_1 f_i / \sum_{i=1}^{n} x_i f_i$
Own ‘price’ elasticity	$\epsilon_i = x_i f_{ii}/f_i$
Elasticity of substitution	$\sigma_{ij} = \frac{\left(f_{ii}/f_i^2 + f_{ij}/f_i f_j - f_{jj}/f_j^2\right)}{\left(1/x_i f_i + 1/x_j f_j\right)}$
Disembodied technological change:	
(1) Rate of technical change	$T = f_t/f$
(2) Acceleration of technical change	$\dot{T} = (f_{tt}/f) - (f_t/f)^2$
(3) Rate of change of marginal products	$\dot{m}_t/m_t = f_{it}/f_i$

Sources: Adapted from Fuss and McFadden (eds), 1978, p. 231

Properties and Form of the Production Functions

The characterization of the input demand and output supply functions depends on the specific properties of the production function. A number of studies have tried to specify these properties and discover more flexible functional forms to accommodate various economic effects often imbedded in the production process. Some economic concepts of interest are listed below. Though the concepts shown in Table 1 are defined in terms of a single-output production function, they can easily be extended to multiple-output production functions. Given the production function $y = f(x, t)$, where x is a vector of inputs and t the index of technological change, it is possible to deduce expressions shown in Table 1 for returns to scale, shares of factors of production, price elasticity and elasticity of substitution, as well as various indices of disembodied technical change. Other effects such as indices of *embodied* technical change can also be derived. By imposing specific restrictions across these effects, different functional forms of the production function can be obtained. Of this array of economic effects, those associated with returns to scale, degree of substitution among inputs and the type and nature of technological change, have received prominent attention in the literature.

These economic effects arise from the inherent nature of the underlying production process, and the specific form of the production function is therefore critical in determining the existence and magnitude of these effects. These properties of the production function – homogeneity, additivity and separability – have played an important role in the derivation of input demand and output supply functions. A homogeneous production function of degree k is defined as:

$$f(\lambda x_1, \ldots, \lambda x_n) = \lambda^k f(x_1, \ldots, x_n); \quad \lambda > 0$$

and a monotonic transformation of a homogeneous production function yields a homothetic production function in $y = g[f(x_1,\ldots,x_n)]$. This family of production functions is characterized by straight-line expansion paths through the origin. Additivity may take the form:

$$f^1(\lambda y_1) + \cdots f^n(\lambda y_n) = f^2(y_1), \ldots, f^n(y_n) = 0 \quad \text{for any} \ \lambda > 0$$

where y_i represents net output of ith commodity, some of which are inputs to the production process. If the function f^i is either homogeneous of some degree, or logarithmic, the additivity condition holds.

Most of the theoretical formulations of the production functions described in the literature implicitly assume that separability conditions prevail. The $f(x)$ is *weakly separable* with respect to partition R when the marginal rate of substitution (MRS) between any two inputs x_i and x_j from any subset $N_{s,}$ $s = 1, \ldots, r$, is independent of the quantities outside N_s (Leontief 1947; Green 1964; Berndt and Christensen 1973) or $\partial(f_i/f_i)/\partial x_k = 0$. *Strong separability*, on the other hand, exists when MRS between any two inputs inside N_s and N_t does not depend on the quantities outside N_s and N_t or $f_i f_{ik} - f_i f_{jk} = 0$.

Functional separability plays an important role in aggregating heterogeneous inputs and outputs, deriving value-added functions and estimating production functions. It also opens up the possibility of consistent multi-stage estimation, which may be the only feasible procedure when large numbers of inputs and outputs are involved in the production activities of highly complex organizations.

A major preoccupation in the literature for empirical estimation of production functions has been to find flexible functional forms. Well-known functions (e.g. the Leontief and Cobb–Douglas production functions) impose restrictions of zero and one, respectively, on the elasticity of substitution, σ while for CES production functions, σ is an arbitrary constant to be estimated. Attempts to relax this stringent requirement have led to the development of the variable elasticity of substitution functions (VES) where σ is dependent on economic variables such as input mix (Liu and Hildebrand 1965; Kadiyala 1972). Efforts to relax the homogeneity property have led to the development of a number of homothetic production functions that make the returns to scale depend on output and/or input mix (Zellner and Revankar 1969; Färe et al. 1978). A major advance has been the formulation of non-homothetic functions by Christensen et al. (1973), who formulated the translog production function, which does not *a priori* impose restrictive constraints such as homotheticity, constancy of σ additivity, and so on.

Technical Progress

Technical progress deals with the process and consequences of shifts in the production function due to the adoption of new techniques which either have a neutral effect on the production process or change the input–output relationships. Neutrality of technical progress can be measured by its effect on *certain* economic variables such as capital–output, output–labour and capital–labour ratios, which should remain invariant under technical change. Several definitions of technical progress have been proposed, such as (1) product-augmenting, (2) labour- or capitalaugmenting, and (3) input-decreasing and factor-augmenting, amongst others (Beckmann et al. 1972). However, the most familiar definitions are the Hicks, Harrod, and Solow forms of technical progress.

Part of technical change can be endogenous and would be determined by the firm to maximize its long-run profit. Technical knowledge is expensive to produce but, once produced, its transmission cost is almost zero, giving rise to the ‘indivisibility’ and ‘inappropriability’ characteristics of inventions. Attempts have been made to incorporate R&D as an input in the neoclassical production and cost functions, to estimate its contributions to the firm’s productivity growth and cost behaviour, and to measure its spillover effects on other firms or industries (Nordhaus 1969; Griliches 1979.) The results indicate substantial private and social rates of return to R&D (Mansfield 1969). Changes in relative prices and output not only affect endogenous technical change but also the rate of factor productivity and the bias of technical change, which will in turn alter the structure of the production process (Jorgenson and Fraumeni 1981).

Duality

A major advance in the economic theory of production has been the dual formulation of production theory (Shephard 1953; Diewert 1974; Fuss and McFadden 1978). The main features of this approach is to recover through indirect functions – that is, by means of a dual representation such as profit or cost functions – the

Production: Neoclassical Theories, Table 2 Comparison of the properties on the transformation function and its dual cost function

Property A on the transformation function $F(y, x)$	Property B on the cost function $C(y, p)$
1 Non-increasing in y	Non-decreasing y
2 Uniformly decreasing in y	Uniformly increasing in y
3 Strongly upper semi-continuous in (y, x)	Strongly lower semi-continuous in (y, p)
4 Strongly lower semi-continuous in (y, x)	Strongly upper semi-continuous in (y, p)
5 Strongly continuous in (y, x)	Strongly continuous in (y, p)
6 Strictly quasi-concave from below in x	Continuously differentiable in positive p
7 Continuously differentiable in positive x	Strictly quasi-concave from below p
8 Twice continuously differentiable strictly differentiably quasi-concave from below in x	Twice continuously differentiable and strictly differentiably quasi-concave from below in p

properties of the underlying production function. The dual approach not only contributes important insights of its own but also offers more immediate empirical applications. A mapping of the characteristics of the transformation function and its dual cost function is indicated in Table 2.

The cost formulation is used extensively in econometric studies. This approach has two main advantages: (1) demand and supply functions can be derived as explicit functions of relative price and output without imposing arbitrary constraints on production patterns required in the traditional methodology; (2) cost and profit functions are computationally simple and permit testing of a wider class of hypotheses by utilizing economic variables (Nadiri 1982).

Dynamic Factor Demand Models

These types of production functions emphasize the intertemporal aspect of the production process by focusing on the movement from one equilibrium state to another. The models incorporate costs of adjustment that are incurred in order to change the level of quasi-fixed inputs, costs which can take two forms. The first type is external: as the firm adjusts its quasifixed factors it must face either a higher purchase price for these factors (Lucas 1967; Gould 1968) or a higher financing cost for the accumulation of these inputs (Steigum 1983). The second type is internal and reflects the fact that firms must make the trade-off between producing current output and diverting some of the resources from current production to accumulate capital for future production (Treadway 1974).

Suppose the firm maximizes its present value:

$$V = \int_0^\infty \{Py - WL - rK - G\dot{K}\}\, \mathrm{e}^{-\rho t}\mathrm{d}t$$

subject to the production function $f(y,L,\dot{K},K) = 0$ and the initial condition $K(0) = K_0$. P is the price of output, y is the level of output, W is the nominal wage, r is the user cost of capital, G is the purchase price of investment, K is a vector of capital inputs, L is labour, and $\dot{K}$ is net investment. $\dot{K}$ is introduced in production on the assumption that firms produce essentially two types of outputs: y, to sell, and $\dot{K}$ the internally accumulated capital which will be used in future production. K is assumed to be neither perfectly fixed nor perfectly variable. Suppose, in addition, that the production function is characterized by the relation $y + C(\dot{K}) - g(K,L) = 0$, where C and g are continuous and the marginal products of f_L and f_K are positive and diminishing.

From the necessary conditions, it follows that for perfectly variable inputs its marginal product must equal its price, while for the quasi-fixed inputs the discounted sum of future net values of its marginal product must equal the sum of the purchase price of investment and the marginal value of real product foregone as a consequence of expansion at the rate $\dot{K}$. The basic problem in this type of model is to deal with expectations about future prices of inputs and outputs. A simple and often-used approach to the problem is to assume static expectations, but that begs the question. Uncertainty about future prices are handled in two ways, either by approximating optimization under uncertainty with certainty equivalence, which requires a quadratic objective function and linear constraint (Hansen and Sargent 1981) or by making adjustment costs a function of the level of the quasi-fixed inputs, which exploits the expectations of future prices that are contained at the quasi-fixed input levels.

There is a fairly large and growing theoretical and empirical literature using the dynamic production function or its duals, dynamic profit or cost functions. The main result of these models is that, because of the existence of adjustment costs, substitution possibilities and technological biases may be limited in the short run, and the effects of prices and tax changes on factor demands may be quite different from their effects in the long run.

Economies of Scale and Scope

An important extension of the theory of the firm has been the production and pricing behaviour of a multi-product firm when economies of scale prevail. To derive the net supply functions, the necessary conditions for equilibrium noted earlier break down when increasing returns or declining long-run average costs prevail. In such cases, monopolistic organization of an industry may offer cost advantages over production by a multiplicity of firms. An interesting and important question is what are the necessary and sufficient conditions for a multi-product firm to be a natural monopoly and for it to be sustainable against entry (Baumol 1977). The condition for natural monopoly is that a cost function be *strictly and globally subadditive* in the set of commodities $C(y^1 + \bullet \bullet + y^m) < C\ (y^1) + \bullet\ \bullet + C(y^m)$, which means that it is always cheaper to have a single firm produce whatever combinations of output is supplied to the market. If the output vectors are restricted to be orthogonal, then the production function exhibits *economies of scope*. The natural monopoly is a sustainable set of products set at prices that do not attract rivals into the industry (Baumol et al. 1977). Even if rivals are attracted, the monopoly may be able to protect itself from entry by changing its prices. But, by definition, only a sustainable vector of prices can prevent entry and yet remain stationary. The conditions necessary for sustainability are (1) the products are weak gross substitutes; (2) the cost function exhibits strictly decreasing ray average costs; and (3) the cost function is also transray convex. Ramsey prices often ensure sustainability under specified circumstances.

See Also

▶ Cobb–Douglas Functions
▶ Cost and Supply Curves
▶ Cost Functions
▶ Humbug Production Function
▶ Joint Production
▶ Supply Functions

Bibliography

Baumol, W.J. 1977. On the proper cost tests for natural monopoly in a multi-product industry. *American Economic Review* 67(5): 809–822.

Baumol, W.J., E.E. Bailey, and R.D. Willig. 1977. Weak invisible hand theorems on the sustainability of multiproduct natural monopoly. *American Economic Review* 67(3): 350–365.

Beckmann, M., R. Sato, and M. Schupack. 1972. Alternative approaches to the estimation of production functions and of technical change. *International Economic Review* 13: 33–52.

Berndt, E.R., and L.R. Christensen. 1973. The internal structure of functional relationships: Separability substitution and aggregation. *Review of Economic Studies* 40(3): 403–410.

Christensen, L.R., D.W. Jorgenson, and L.J. Lau. 1973. Transcendental logarithmic production frontiers. *Review of Economics and Statistics* 55(1): 28–45.

Diewert, W.E. 1974. Applications of duality theory. In *Frontiers of quantitative economics*, vol. II, ed. M.D. Intriligator and D.A. Kendrick. Amsterdam: North-Holland.

Färe, R., L. Jansson, and C.A. Knox Lovell. 1978. On ray-homothetic production functions. In *The importance of technology and the permanence of structure in industrial growth*, vol. II, ed. B. Carlsson, G. Elliasson, and M.I. Nadiri, 228–237. Stockholm: Industrial Institute for Economic and Social Research.

Fuss, M., and D. McFadden (eds.). 1978. *Production economics: A dual approach to theory and application*, vol. I. Amsterdam: North-Holland.

Gould, J.P. 1968. Adjustment costs in the theory of investment of the firm. *Review of Economic Studies* 35: 47–55.

Green, H.A.J. 1964. *Aggregation in economic analysis: An introductory survey*. Princeton: Princeton University Press.

Griliches, Z. 1979. Issues in assessing the contribution of research and development to productivity growth. *Bell Journal of Economics* 10(1): 92–116.

Hansen, L.P., and T.J. Sargent. 1981. Linear rational expectations models for dynamically interrelated variables. In *Rational expectations and econometric practice*, vol. I, ed. R.E. Lucas and T.J. Sargent, 127–156. Minneapolis: University of Minnesota Press.

Jorgenson, D.W., and B.M. Fraumeni. 1981. Relative prices and technical change. In *Modeling and measuring natural resource substitution*, ed. E.R. Berndt and B.C. Fields, 17–47. Cambridge, MA: MIT Press.

Kadiyala, K.R. 1972. Production functions and elasticity of substitution. *Southern Economic Journal* 38(3): 281–284.

Leontief, W.W. 1947. Introduction to a theory of the internal structure of functional relationships. *Econometrica* 15: 361–373.

Liu, T.C., and G.H. Hildebrand. 1965. *Manufacturing production functions in the United States, 1957*. Ithaca: Cornell University Press.

Lucas, R.E. 1967. Adjustment costs and the theory of supply. *Journal of Political Economy* 74(4): 321–334.

Mansfield, E. 1969. Industrial research and development: characteristics, costs and diffusion of results. *American Economic Review* 59(2): 65–71.

Nadiri, M.I. 1982. Producers theory. In *Handbook of Mathematical economics*, vol. II, ed. K.J. Arrow and M.D. Intriligator, 431–490. Amsterdam: North-Holland.

Nordhaus, W. 1969. *Invention, growth and welfare: A theoretical treatment of technological change*. Cambridge, MA: MIT Press.

Shephard, R.W. 1953. *Cost and production functions*. Princeton: Princeton University Press.

Steigum, E. 1983. A financial theory of investment behavior. *Econometrica* 51(3): 637–645.

Treadway, A. 1974. The rational multivariate flexible accelerator. *Journal of Economic Theory* 7(1): 17–39.

Zellner, A., and N. Revankar. 1969. Generalized production functions. *Review of Economic Studies* 36(2): 241–250.

Productive and Unproductive Consumption

Mark Blaug

The terms ‘productive’ and ‘unproductive’ consumption were introduced into economics by John Stuart Mill in Book I, chapter 3, of his *Principles of Political Economy* (1848), although the distinction (without the terminology) first appeared in the closing pages of the new chapter on machinery in the third edition of Ricardo’s *Principles* (1821), where he argued that workers have a personal interest in the pattern of luxury spending by the rich because spending on ‘menial servants’ increases the demand for labour by more than the equivalent amount of spending on physical goods. Mill’s distinction between the two kinds of consumption is a direct application of Smith’s distinction between ‘productive’ and ‘unproductive’ labour; according to Mill, the only productive consumers are productive labourers, but not all consumption by productive labourers is productive consumption: ‘that alone is productive consumption’, Mill observes, ‘which goes to maintain and increase the productive powers of the community’.

The basic idea behind the distinction between productive and unproductive consumption goes back to the writings of the physiocrats. It is the notion that a certain quantity of the consumer goods produced in an economy (i.e. wage goods) enters as necessary inputs into the production of manpower itself in the household sector. Productive consumption is simply an input necessary to maintain human capital intact. If wages are at subsistence, the whole of the wages bill is required for productive consumption. Mill concedes, however, that workers do consume some ‘luxuries’, and in that sense a portion of wages is consumed unproductively. The fact remains that consistent classical income accounting consistent with the subsistence theory of wages implies deducting all productive consumption from the gross national product to arrive at the true net national product, which consists simply of profits plus rent; this net product is created entirely by productive labour and is spent entirely on investment and *true* consumption goods, that is, non-wage goods. The logic of this argument is impeccable, although the statistical difficulty of segregating wages into its productive and unproductive components might be unsurmountable. The point is, however, that only a society bent on maximizing capital accumulation, come what may to current living standards, would want to adopt this kind of accounting. But

Mill, unlike Adam Smith, was not at all sure that a higher rate of economic growth was really desirable. Thus the only practical use that Mill made of the distinction between productive and unproductive consumption was to convert Ricardo's thesis that the volume of employment depends on the pattern of unproductive consumption into the paradoxical proposition that 'demand for commodities is not demand for labour'.

The dozen or so pages explaining this proposition in Mill's *Principles* (Book I, ch. 5, para. 3) are among the most tortuous ever to have been penned by a great economist, and the secondary literature commenting on Mill's discussion is almost entirely negative (Thompson 1975). Nevertheless, it is possible to catch the drift of Mill's meaning, which was that the volume of employment in an economy is a direct function of the rate of capital accumulation, so that consumers' demand, while it clearly determines the allocation of labour between different industries, influences total employment only at one remove. Since the decision whether the proceeds of sales will be used to reconstitute the 'wages fund' for a new bout of production rests with employers, demand for commodities is not *necessarily* demand for labour. Having made the decision to save a certain portion of his income, the only way in which an individual can directly influence the demand for labour is by substituting labour services for commodities in his own consumption. This is Ricardo's old argument that the interest of labour is best served by the most labour-intensive kind of spending on personal consumption.

Unfortunately, Ricardo had applied his argument to a situation in which some labour has become unemployed owing to the sudden introduction of labour saving machinery, from which it follows that demand for commodities *is* demand for labour. Mill, however, seems to be assuming full employment by affirming that an increased demand for labour in one industry must draw labour out of another. In that case it seems to follow tautologically that an increased demand for consumer goods cannot increase the demand for labour. But Mill's object was to argue the stronger thesis that the demand for labour will in fact fall off under full employment when resources are shifted into the manufacture of additional consumer goods: an increase in consumption means a decrease in investment, and investment under the wages fund doctrine can only mean advancing more wage goods to labour in subsequent periods. Given the rigid discontinuity of production implied in the wages fund doctrine, it is perfectly true that an increase in aggregate consumption demand under full employment impairs the wages fund and so leads to a decline in the amount of employment demanded at any given wage rate. In short, Mill's proposition that 'demand for commodities is not demand for labour' holds only insofar as the wages fund doctrine itself.

With the demise of the wages fund doctrine, the distinction between productive and unproductive consumption disappeared from the literature of economics. Marshall (1890, p. 67) included the distinction in his *Principles* but made absolutely no use of it and added characteristically in a footnote: 'All the distinctions in which the word "productive" is used are very thin and have a certain air of unreality. It would be hardly worthwhile to introduce them now: but they have a long history; and it is probably better that they should dwindle gradually out of use, rather than be suddenly discarded.' The reason for the disappearance of the Millian distinction is obviously connected with the declining interest after Mill in the fundamental question raised by any such distinction; namely, what is it that determines the volume of employment? Thus Marshall and many other neoclassical economists of the post-1870 period retained an interest in the distinction between different types of consumption as between) 'necessities', 'convenience' and 'luxuries' in terms of their different price elasticities of demand rather than the volume of employment generated by the pattern of expenditure on different categories of consumer goods.

See Also

- ▶ British Classical Economics
- ▶ Smith, Adam (1723–1790)

Bibliography

Marshall, A. 1890. *Principles of economics,* vol. 1, 9th variorum ed, ed. C.W. Guillebaud. London: Macmillan, 1961.

Thompson, J.H. 1975. Mill's fourth fundamental proposition: A paradox revisited. *History of Political Economy* 7(2): 174–192.

Productive and Unproductive Labour

Guido Montani

According to Schumpeter the debate on productive and unproductive labour was nothing but a 'dusty museum piece' (Schumpeter 1954, p. 628). And indeed, after the achievement of marginal utility theory, there was no need to distinguish between productive and unproductive labour, because all labour producing 'useful and scarce' things was to be considered as productive. So the meaning of 'productive' covers the whole field of economic goods. Albeit cautiously, Marshall suggested dropping this kind of terminology. 'Whenever we use the word Productive of itself [says Marshall] it is to be understood to mean *productive of the means of production, and of durable sources of enjoyment*. But it is a slippery term, and should not be used where precision is needed. If ever we want to use it in a different sense, we must say so: for instance we may speak of labour as *productive of necessaries, etc*.' (Marshall 1890, p. 56).

Nevertheless, in classical economics the distinction between the two kinds of labour seems to be a very useful one, since it makes it possible to know whether a certain expenditure of money can engender a new income (in which case it is to be considered an investment) or whether it engenders only enjoyment or waste (in which case it is to be considered consumption). At the very beginning of the *Wealth of Nations* Adam Smith says that the nation will be better or worse off according to 'the proportion between the number of those who are employed in useful labour, and that of those who are not so employed' (Smith 1776, p. 1). This distinction raised innumerable disputes during the last century, but, in a different way, the problem is not at all avoidable in contemporary economics. Take for instance the question of including civil service expenses in national accounting or not. Should we consider expenses for higher education as consumption, as investment or as a demonstration of affluence? And the defence budget? Is the national income increased or decreased when more nuclear weapons are produced and stored in underground silos? There is no clear-cut reply. As Marshall warned, if we take a certain expense as consumption or investment according to its capacity to 'produce' only utility (enjoyment) or new income as a means of production, a nuclear weapon is neither a consumption nor an investment good.

Here the difficulty is raised by the fact that a certain activity is considered 'useful' or not for political reasons, social welfare policies, etc., that is, criteria not based on market evaluations. Therefore, we can say that a certain State expenditure, for instance, in the defence or education sector, is *indirectly useful* to the production of national income, since without State administration the very existence of a free market is impossible.

Even if we tie ourselves to market produced commodities some problems still arise. Take, for instance, Sraffa's modern reconstruction of the classical theory of value and distribution.

Sraffa does not take into consideration the distinction between productive and unproductive labour. But when he comes to examine the case of production with surplus, he states that) 'one effect of the emergence of a surplus' is that 'there is room for a new class of "luxury" products which are not used, whether as instruments of production or as articles of subsistence, in the production of others. These products have no part in the determination of the system. Their role is purely passive' (Sraffa 1960, p. 7). Therefore, if we define all the labour employed in sectors which contribute directly or indirectly to the production of the social surplus as productive, the labour employed in the 'luxury' sectors ought to be considered as unproductive. If we eliminate, during a certain year, the activities relating to 'luxury' productions 'the price-relations of the other products and the rate of profits would remain unaffected'. The following year, the production cycle can start anew on the same scale and the same quantities of 'non-luxury' (basic) commodities can be produced.

Nevertheless, there is no perfect match between the notion of luxury and unproductive sectors. Take, for example, commerce, which was considered by Marx as unproductive (see *Capital*, vol. II). The whole economy is divided into two sectors: industry, where commodities are produced and brought at the end of the year to the wholesale market, and commerce, where the surplus is sold on the retail market during the year. Therefore, the commerce sector does not produce any physical new product. Its function is to transfer the commodities required by consumers to other places and to store them up during the year. The rate of profits, if wages are given, is determined only inside the industrial sector. Merchants are able to get a rate of profits equal to the industrial one if they sell commodities at a retail price higher than the price paid to industries on the wholesale market.

From a formal point of view, commerce should be considered as a luxury sector, because all commodities sold on the retail market do not enter, directly or indirectly, into the industrial sector as means of production. But two more observations should be considered: (a) commodities are bought on the retail market mostly by workers and therefore, as wage goods, they enter industrial production; (b) the existence of a commercial sector side by side with the industrial one means that a better division of labour is possible (as compared with the case where commerce does not exist) and that consumers can buy commodities at a cheaper price. If it were not so, consumers would find it convenient to go directly to the factory to buy commodities. Therefore, we must say that the commercial sector is 'productive', because, if we do away with it, it would be impossible for the economy to reproduce the same quantities of commodities the following year with the given quantity of the work-force previously employed (of course, no technological change is considered here).

So far, we have worked with a concept of productive labour in a simple reproduction economy, that is, an economy which reproduces itself every year on the same scale. This concept does not change even if we consider the case of expanded reproduction, that is, accumulation. Accumulation means only an enlargement of the productive forces already employed, and not a qualitative change in the role played by every sector of the economy in relation to other sectors. In fact, one of the main problems of an expanding economy is the study of the proportion which should exist between sectors, on the one hand, and between production and income, on the other, in order to assure steady and balanced growth.

The case of development is very different. By development we mean the transition from a certain 'mode of production' (to use Marxist terminology) to another one. This approach was typical of classical economists and their followers, but it seems to have been forgotten in contemporary economic analysis. For instance, Adam Smith refers over and over again to 'that early and rude state of society' which preceded the capitalist development and the industrial revolution. It is worthwhile here remembering that a vast literature flourished during the 18th century on the 'four stages of progress' which mankind passed through (Meek 1976).

In the history of economic thought the best treatment of the problem of development is in List's *National System of Political Economy* (1841), where a sketch of a theory of productive forces (or productive powers) can be found. List protested against the tendency to reduce political economy to the theory of value. List stated clearly 'that an independent *theory of the "productive power"* must be considered by the side of a *"theory of values"* in order to explain the economical phenomena' (List 1841, p. 137). Smith's and Say's concept of productive labour, according to List, is very limited because it refers only to the creation of exchangeable values. On the contrary, we should consider all the expenses which produce *productive powers* as productive.

> The errors and contradictions of the prevailing school [says List] can be easily corrected from the standpoint of *the theory of the productive powers*. Certainly those who fatten pigs or prepare pills are productive, but the instructors of youths and of

> adults, virtuosos, musicians, physicians, judges, and administrators, are productive in a much higher degree. The former *produce values of exchange*, and the latter *productive powers* (List 1841, p. 143).

List's concept of productive powers is very useful in contemporary economic analysis. Let us consider a typical problem of underdevelopment. In an agricultural society, public expenses for primary education should be considered as unproductive if our point of view is limited to a selfreproduction economy. But if our aim is the transition towards an industrial society, primary education should be considered as a prerequisite, since industrial production needs a skilful workforce, businessmen, a service sector, etc. In a few words: we need to create productive powers in order to go beyond the agricultural mode of production.

But List's theory of productive powers is very helpful for a second reason: it allows us to consider the *State* as a productive power (or force). In fact, List aimed at the political unification of Germany (a Customs Union) in order to create conditions favourable to German entrepreneurship, which was choked in his day by stronger English industrial competition. In our century, one of the most striking phenomena is the transition of the main industrialized countries (both capitalist and socialist) towards a new mode of production, which should be called 'scientific', given the role played by science as a productive force. But since science, and typically 'big science' which requires huge public investment, can be organized efficiently only by state research programmes, it seems correct to consider all public expenses nowadays devoted to scientific research as 'productive'.

See Also

- ▶ British Classical Economics
- ▶ Labour Theory of Value

Bibliography

List, F. 1841/1966. *The national system of political economy*. New York: A.M. Kelley.

Marshall, A. 1890/1920. *Principles of economics*, 8th edn. London: Macmillan. Reprinted, 1972.

Meek, R.L. 1976. *Social science and the ignoble savage*. Cambridge: Cambridge University Press.

Schumpeter, J.A. 1954. *History of economic analysis*. New York: Oxford University Press.

Smith, A. 1776/1964. *An inquiry into the nature and causes of the wealth of nations*. London: Everyman's Library.

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

Productivity: Measurement Problems

Zvi Griliches

Productivity is a ratio of some measure of output to some index of input use. The meaning and quality of such a measure depends on the definition and quality of its ingredients and on the particular formula and the associated weights used to aggregate the various components into one output or input index. Economists tend to think of productivity as measuring the current state of technology used in producing the goods and services of an economy (or industry or firm), and want to interpret the changes in such a measure as reflecting) 'technological change', shifts in the production possibilities frontier. For this purpose, it is usual to focus on one or another version of 'multi-factor productivity', where the list of inputs considered extends beyond labour and includes also measures of capital services and also, occasionally the use of materials and other inputs.

Measuring technological change is of interest because, in a sense, it defines our wealth and puts limits on what we can accomplish. Our wealth at any time can be thought of as consisting of three parts: (a) quantities of various resources available to the economy – labour, capital, land, minerals, etc.; (b) the organizational arrangements for using these resources in production; and (c) the currently known ways of converting such resources into outputs desired by the economy. Each one of these can potentially increase, improve, or decline and deteriorate. The range of things covered under the last broad concept – 'technology' – can be thought of as consisting of both the average set of recipes for doing things, embedded in the current range of equipment and in the training of the existing work-force, and the currently known *best* way of doing things. 'Technological change' as measured by productivity type measures, even after allowances have been made for the growth in other resources, for economies of scale, and for errors in measurement, reflects changes in the 'average' technology used by an industry or economy. 'Average' technology can improve because more firms have shifted to the best (or better) technology (this is called diffusion of technology), because some of the poorer firms have gone out of business even though the others have not become any better, and because the 'best' technology itself has improved, the technological frontier having been expanded by science, organized research, learning by doing, and plain serendipity. In the long run, average technology can keep improving only if the best technology is also improving. Ultimately, therefore, since our ability to accumulate additional conventional resources, such as capital or mineral resources, may be limited, the growth of the economy and of per capita income and wealth depends on the rate at which technological knowledge is expanding and on our ability to affect this rate by changing the amount of resources we devote to science and technology and on the institutions that we devise for using them effectively.

The point here is that we are interested in 'productivity' and what is happening to it primarily because we believe that it may tell us something about the ultimately more fundamental process of technological change. But conventional measures of productivity are only a distant and murky reflection of it. Changes in such indexes, especially sharp short-run fluctuations, should be interpreted with great care since they may have little to do with technological change proper.

The list of potential problems is long and overlapping:

1. Coverage issues, definition of the borders of a sector and of the relevant concept of 'output' for it. For example, is illegal activity included? Are pollution damages counted against the 'output' of an industry?
2. The measurement of 'real' output over time. Price) 'deflation' and quality-change problems.
3. The measurement of inputs over time. The changing skill-mix of the labour force, quality change in the machinery and equipment used and the changing utilization of the labour force and of the existing capital stock.
4. The list of inputs to be included in the total input concept. The treatment of research and development and of public infrastructure expenditures.
5. Missing data on hours worked by people and machines and/or specific input use by various industries.
6. Getting the right 'weights'. The divergence of market prices from 'shadow prices' and the impact of various disequilibria.
7. Formula differences. Index number formulae and the unknown shape of the underlying production possibilities frontier. Gross versus net concepts and other variants.
8. The consequences of aggregation over heterogeneous individuals and industries.

The last two topics are treated at length in separate entries (*see* AGGREGATION PROBLEM and INDEX NUMBERS) and will not be considered further here. It may be useful to outline briefly an algebraic–taxonomic framework for organizing the discussion about this range of topics. The conventional measure of residual technical change (TFP) in an industry can be written as

$$\lambda = y - sk - (1 - s)n$$

where *y, k*, and *n* are percentage rates of growth in output, capital, and labour respectively, *s* is the share of capital in total factor payments, and the relevant notion of capital corresponds to an aggregate of actual machine hours weighted by their respective base period (equilibrium) rentals. This procedure assumes that all the variables are measured correctly, that all the relevant variables are included, and that factor prices represent adequately the marginal productivities of the respective inputs. The last assumption is equivalent to the assumption of competitive equilibrium and constant returns to scale. To analyse λ the 'unexplained' part of output growth, it is useful to rewrite it in terms of a more general underlying production function and the 'correct' set of inputs:

$$\begin{aligned}\lambda = {} & s(k^* - k) + (1 - s)(n^* - n) \\ & + (s^* - s)(k^* - n^*) \\ & + h[sk^* + (1 - s^*)n^* - f] + \alpha_z z + u + t\end{aligned}$$

where the starred magnitudes are rates of growth of the correctly measured inputs; $s^* = \alpha_k/(\alpha_k + \alpha_n) = \alpha_k/(1 + h)$ where the α ' s are the true elasticities of output with respect to the specific inputs and $h = \alpha_k + \alpha_n - 1$ is a measure of economies of scale with respect to the conventional inputs k and $n; f$ is the percentage rate of growth in the number of establishments (plants) in the industry; and z is the rate of growth of inputs which affect output but are not included in the standard accounting system. These could be services from the accumulated stock of past private research and development expenditures, or services from the cumulated value of public (external) investments in research and extension in agriculture and other industries, or measurable disturbances such as weather or earthquakes. The last terms, u and t, represent respectively errors in the measurement of output growth and the 'true' rate of growth in the average level of technology. The α coefficients, h, and $s*$ need not be constants. If they are, we have the Cobb–Douglas case. The whole framework could be complicated and generalized by adding square terms in rates of growth as approximations to CES or translog type production functions.

The various terms in this formula can be interpreted as follows: The first term is the effect of the rate of growth in the measurement error of the conventional capital input measure on the estimated 'residual'. The second term reflects errors in the measurement and definition of labour input. The third term reflects errors in assessing the relative contribution of labour and capital to output growth. It would be zero if factor shares were in fact proportional to their respective production function elasticities. The fourth term is the economies of scale term. It would be zero if there are no underlying economies of scale in production ($h = 0$) or if the rate growth in the number of new firms (plants) just equalled the growth in total (weighted) input. The fifth term reflects the contribution of omitted inputs (private or public). The sixth term (u) represents various unspecified errors in the measurement of output growth, while the last term t is the 'pure' residual term – the amount of output growth not accounted for by this expanded list of possible sources.

Turning to a discussion of the potential content of some of these empty boxes, let us start with the penultimate term (u), errors in the measurement of output growth, before turning to the consideration of the input side. As far as the definition and measurement of output is concerned there are serious definitional problems at the aggregate national level about the borders of economic activity (e.g. home production and the investment value of children) and where one should draw the line between final and intermediate consumption activity (e.g. what fraction of education and health expenditures can be thought of as final rather than intermediate 'good' or 'bad'?). There are also difficult measurement problems associated with the existence of the underground economy and the poor coverage of some of the major service sectors. The most serious problem is probably in the measurement of 'real' output in 'constant prices' (at the national or industry level) and the associated growth measures. Since many output measures are derived by dividing ('deflating') current value totals by some price index, the quality of these measures is intimately connected to the quality of the available price data.

Because of this, it is impossible to treat errors of measurement at the aggregate level as independent across price and) 'quantity' measures.

The available price data, even when they are a good indicator of what they purport to measure, may still be inadequate for the task of deflation. For productivity comparisons and for production function estimation the observed prices are supposed to reflect the relevant marginal costs and revenues in a competitive equilibrium. But this is unlikely to be the case in sectors where output or prices are controlled, regulated, subsidized, and sold under various multi-part tariffs. Because the price data are usually based on the pricing of a few selected items in particular markets, they may not correspond well to the average realized price for the industry as a whole during a particular period, both because) 'easily priced' items may not be representative of the average price movements in the industry as a whole and because many transactions are made with a lag, based on long-term contracts. There are also problems associated with getting accurate transaction prices but it is the continued change in the available set of commodities that creates major difficulty: the 'quality change' problem.

'Quality change' is actually a special version of the more general comparability problem, the possibility that similarly named items are not really similar, either across time or individuals. In many cases the source of similarly sounding items is quite different: Employment data may be collected from plants (establishments), companies or households. The answer to the same question may then have a different meaning, depending on the source.

The common notion of quality change relates to the fact that many commodities are changing over time and that often it is impossible to construct appropriate pricing comparisons because the same varieties are not available at different times and in different places. Conceptually one might be able to get around this problem by assuming that the many different varieties of a commodity differ only along a smaller number of relevant dimensions (characteristics, specifications), estimate the price–characteristics relationship econometrically and use the resulting estimates to impute a price to the missing model or variety in the relevant comparison period. This has become known as the 'hedonic' approach to price measurement. The data requirements for the application of this approach are quite severe and there are very few official price indexes which incorporate it into their construction procedures.

While there have been significant improvements in data collection and processing procedures over time, it is fair to note that much still remains to be done. In the US GNP deflation procedures, until recently the price of computers had been kept constant since the early 1960s, for lack of agreement on what to do about it, resulting in a significant underestimate in the growth of real GNP during the last two decades. There are more such horror stories to be told but the point here is not that a particular price index is biased in one or another direction, rather that one cannot take a particular published price or 'real output' index series and interpret it as measuring adequately the underlying notion of price or output change for a well-specified, unchanging commodity or service being transacted under identical conditions and terms in different periods. The particular time series may indeed be quite a good measure, or at least better than the available alternatives, but each case requires a serious examination of whether the actual procedures used to generate the series do lead to a variable that is close enough to the concept envisioned by the model to be estimated, or by the theory under test.

The issues discussed above affect also the construction and use of various 'capital' measures in the analysis of productivity growth. Besides the usual theoretical issues of aggregation connected with the 'existence' of any unambiguous capital concept, the available measures suffer from potential quality change problems, since they are usually based on some cumulated function of past investment expenditures deflated by some combination of available price indexes. In addition, they are also based on rather arbitrary assumptions about the time pattern of both the survival of the machines themselves and the deterioration in their flows of services. The available information on the reasonableness of such assumptions is very sparse, ancient and flimsy. In some contexts it is possible to estimate the appropriate pattern from the data rather than impose them a priori, but very little of that has been incorporated into the conventional estimates.

From the production function point of view, what is needed is a measure of the flow of services of capital in constant prices. A central problem in constructing such a measure is what to assume about how the services of a given machine behave as it ages. It is usually assumed that they decline rapidly with age, as evidenced by the undoubted fact that the price of a used machine falls rapidly with its age. But the value of an old machine declines because its expected worklife is falling, because new and better machines are becoming available, and because the quality of the services it renders deteriorates with its age. Only the last factor is a legitimate deduction to be made from a service-oriented measure of capital. Admittedly, there is less life left in the old machine; but that does not imply that its current product is any the worse for that. Of course, the availability of better machines will result in capital loss for the owners of the old machine; but that does not make the old one any worse, only the new one better.

Even if capital services were in fact proportional to their stock measure (if, say, they did not deteriorate with age – the one-hoss shay – or declined exponentially) this property would not continue to hold if one added together machines with different lengths of life. A $100 machine that will last five years will have roughly twice as large an *annual* flow of services (in dollars) as another $100 machine whose expected length of life is ten years. Thus, the shorter the life expectancy, the higher is the ratio of services to stock. In manufacturing we can identify two major components of capital formation: equipment and structures. Structures have a much longer life than equipment and hence should be given a lower weight in compiling an index of capital *services*. Since the stock of equipment has been growing more rapidly, this adjustment makes a substantial difference to our measurement of the growth in the total level of capital services.

The other major problem with the measurement of capital is the lack of good data on its rate of utilization. Ideally, the relevant measure of capital services is one that is close to the number of machine-hours worked, weighted by their respective rental rates. Business cycle fluctuations and errors in forecasting demand result in large fluctuations in the rate of utilization of installed capacity. Such fluctuations are relevant for ‘efficiency’ measures of productivity, since they tell us how well we are utilizing our existing resources but are misleading as far as ‘technological change’ measures are concerned. It is very difficult, however, to find or design good capital utilization measures. Most of the existing measures are based on deviations of output from trend or previous peaks and contain essentially no independent information on this question.

Similar issues arise also with respect to the third term in this formula, the measurement of the contribution of labour input and associated variables: hours of work, unemployment, and wage rates; both at the macro and micro levels. At the macro level the questions turn on the appropriate weighting to be given to different types of labour: young–old, male–female, educated versus uneducated, and so forth. The direct answer here as elsewhere is that they should be weighted by their appropriate marginal prices; but whether the observed prices actually reflect correctly the underlying differences in their respective marginal productivities is one of the more hotly debated topics in labour economics.

It is also difficult to find relevant labour quantities and prices. The usual data sources report average annual, weekly, or hourly earnings which do not represent adequately either the marginal cost to the employer or the marginal return to a worker of an additional hour of work. Both are affected by the existence of overtime premia, fringe benefits, training costs, and transportation costs. Only recently has an employment cost index been developed in the United States. From an individual worker’s point of view the existence of non-proportional tax schedules introduces another source of discrepancy between the observed wage rate and the unobserved marginal after-tax net return from working. In addition, there are serious problems associated with the measurement of the number of hours worked rather than just hours paid for.

The fourth term is a wrong relative weights or disequilibrium term. It comes into play when the underlying elasticities of the production relation are not well approximated by the respective factor shares and the different inputs are not growing at the same rate. It could be of importance in agriculture in some periods, if capital use is growing while labour use is declining, while at the same time the observed relative prices underestimate the discrepancy in their marginal products.

The fifth term reflects the impact of economies of scale. It depends both on the degree of homogeneity of the production function and the amount of growth in average plant size that occurred during the period in question.

The sixth term reflects the contribution of ‘unconventional’ inputs, inputs that should have been, but have not been included in the standard accounting framework, either through error or because they are ‘external’ to the units being analysed (e.g. the resources used in the construction and maintenance of airports and air-traffic control systems as an input into the production function of the air-transport industry).

Because some of these inputs are not marketed directly and hence have no price attached to them which could be used to approximate their marginal product, and because the previous two terms (wrong weights and economies of scale) also imply a divergence of true productivity from observed market data, one cannot learn about them from an examination of conventional accounting data. Econometric estimation of production functions is required to assess the importance of economies of scale, to test the divergence of estimated coefficients from observed factor shares, to estimate the contribution of excluded and public inputs such as research and development expenditures, and to validate suggested quality adjustments for particular inputs (such as the measurement of the quality of labour by wage-weighted education per man indexes). Production (or cost) function based approaches to the measurement of technical change raise many difficult problems of their own, but they are the only way of testing our notions of what is the ‘right’ way of measuring certain inputs and their contribution to overall output growth.

It is hard to single out one factor or measurement problem as more important than another. Their import differs across countries, industries, and time. In the recent past, the major problems (or villains) have been the measurement of real output and real input growth, i.e. the correct measurement of prices and the associated adjustment for quality change and, in the context of shorter-run comparisons, the treatment of capacity utilization fluctuations. Over a longer historical view, the growing quality of the labour force and the discovery (and exhaustion) of new resource pools have probably been both the major source of true productivity growth and the major source of its mismeasurement, or at least misinterpretation.

See Also

- ▶ Growth Accounting
- ▶ Technical Change
- ▶ Total Factor Productivity

Bibliography

Denison, E.F. 1969. Some major issues in productivity analysis: An examination of estimates by Jorgenson and Griliches. *Survey of Current Business*, Part II, May, including reply by Jorgenson and Griliches.

Denison, E.F. 1979. *Accounting for slower economic growth: The U.S. in the 1970's*. Washington, DC: Brookings.

Gollop, F.M., and D.W. Jorgenson. 1980. U.S. productivity growth by industry, 1947–1973. In *New developments in productivity: Measurement and analysis*, Studies in income and wealth, vol. 44, ed. J.W. Kendrick and B.N. Vaccara. Chicago: University of Chicago Press for the National Bureau of Economic Research.

Griliches, Z. 1970. Notes on the role of education in production functions and growth accounting. In *Education, income and human capital*, Studies in income and wealth, vol. 35, ed. W.L. Hansen. New York: NBER.

Griliches, Z. (ed.). 1971. *Price indexes and quality change*. Cambridge, MA: Harvard University Press.

Griliches, Z. 1979. Issues in assessing the contribution of research and development to productivity growth. *Bell Journal of Economics* 10(1): 92–116.

Griliches, Z. 1985. Economic data issues. In *Handbook of econometrics*, vol. III, ed. Z. Griliches and M. Intriligator. Amsterdam: North-Holland.

Jorgenson, D., and Z. Griliches. 1967. The explanation of productivity change. *Review of Economic Studies* 34(3): 249–283.

Produit Net

G. Vaggi

The French Physiocrats defined the net product, *produit net*, as the difference between the gross output of agriculture and the overall expenses involved in production (Mirabeau 1764, vol. 1, p. 337; Quesnay 1768, p. 979). The *produit net*, or revenue as they often called it, is that part of annual output which is left over after the deduction of the means of production which have been consumed. It is a great merit of Quesnay to have provided the first clear distinction between gross and net product.

The Physiocratic concept of *produit net* has two characteristics; it can be expressed in physical terms as a surplus of agricultural output over its inputs. But the Physiocrats measured the net product also in terms of value, as the difference between the value of agricultural product and the overall value of the expenses incurred in its production. It was a peculiar opinion of the Physiocrats that only agriculture and the other activities directly linked to nature could give rise to a net revenue over costs. According to them the value of industrial output was just equal to that of its costs of production. Of course, from the works of Turgot and Smith onwards industry was considered productive as well as agriculture, and the net product then included capitalists' profits and not only rent (Smith 1776, vol. 1, pp. 330 ff.).

The contention that only the primary sector gives rise to a *produit net* is the cornerstone of Physiocratic economic policy. The net product is the only part of the output which is freely disposable without jeopardizing further production (Quesnay 1766a, p. 869). Thus the concept of *produit net* provides the basis for the Physiocratic theory of taxation; only a single tax on rent does not damage agricultural production, and hence the whole economy. Commercial policies must favour the profitable sale of the products of land in order to sustain their prices and the net product. For the same reason the landlords must spend most of their revenue in the purchase of agricultural goods rather than in that of manufactured commodities.

The Physiocratic notion of *produit net* is important because it originates that stream of thought that since Marx has been called 'theories of surplus' (Marx 1963, vol. 1, p. 44). This approach, which is based on the separation between surplus and capital, is still influential on present-day economic theory (see Sraffa 1960; Pasinetti 1977; Leontief 1941). Quesnay's distinction between productive and sterile activities is based on the criterion of whether or not they yield a *produit net*, a point of view which has also been accepted by Smith. Ricardo and Marx too distinguished productive and unproductive labour.

In Physiocracy the existence of surplus in agriculture was sometimes justified in terms of the benevolence of nature (see Weulersse 1910, vol. II, p. 112–115). But Quesnay defended his view of the exclusive productivity of agriculture with much stronger arguments than the idea of a 'gift of nature'. He linked the existence of a *produit net* to the availability of a large stock of advances in agricultural production, and to the favourable market conditions for the products of French soil (see Quesnay 1767, pp. 960–961; Quesnay 1766b, p. 911; Meek 1962, p. 387). The fact that only agriculture yielded a *produit net* led the Physiocrats to identify the largest part of the net product with the rent of the landlords. But part of the annual revenue of cultivation also goes to the King as taxes.

See Also

▶ Net Product

Bibliography

Leontief, W.W. 1941. *The structure of American economy 1919–1929*. Cambridge: Harvard University Press.

Marx, K. 1963. *Theories of surplus value*, vol. 1. London: Lawrence & Wishart.

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

Mirabeau, V.R. 1764. *Philosophie rurale, ou economie générale et politique de l'agriculture*. Amsterdam:

Libraries Associées. Reprinted, Aalen: Scientia Verlag, 1972.

Pasinetti, L.L. 1977. *Lectures on the theory of production*. London: Macmillan.

Quesnay, F. 1766a. Premier problème économique. In INED, *François Quesnay et la physiocratie*, vol. I. Paris, 1958.

Quesnay F. 1766b. Sur les travaux des artisans. In INED, vol. II, 1958.

Quesnay, F. 1767. Maximes générales du gouvernement économique d'un royaume agricole. In INED, vol. II, 1958.

Quesnay, F. 1768. Second problème économique. In INED, vol. II, 1958.

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

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

Weulersse, G. 1910. *Le mouvement physiocratique en France (de 1756 à 1770)*, 2 vols. Paris: Alcan.

Profit and Profit Theory

Meghnad Desai

Keywords

Abstinence theory of profits; Aggregate production function; Arrow–Debreu model of general equilibrium; Capital accumulation; Capital theory; Clark, J.B.; Cobb–Douglas functions; Constant and variable capital; Contingent commodities; Depreciation; Division of labour; Entrepreneurship; Exchange value and use value; Exploitation; Falling rate of profit; Innovation; Interest; Investment rate; Joint production; Kaldor; Kalecki, M.; Keynes, J. M.; Kondratieff cycles; Labour power; Lamberton, D. M.; Luxemburg, R.; Marginal productivity theory; Marx, K. H.; Organic composition of capital; Physiocracy; Population growth; Produit net; Profit and profit theory; Quasirent; Rent; Risk; Saving and investment; Schumpeter, J. A.; Shackle, G. L. S.; Sraffa, P.; Surplus; Surplus value; Taylorism; Technical progress; Technology; Uncertainty; Use value; Variable capital; von Neumann, J

JEL Classifications

D2

A theory of profit should address itself to at least three questions – about the size (volume) of profit, its share in total income and about the rate of profit on capital invested. Each of these three issues (size, share, rate, hereafter) can be examined at three separate levels of aggregation, the firm, the industry or the economy.

Theories of profits of size, share or rate can be classified as to whether they treat profits as an equilibrium or as a disequilibrium phenomenon, and of course as to whether the equilibrium is a static or a dynamic one. Theories then deal with the role of profits, that is, the effects of the size, share or rate on other economic variables, and the source of profits, that is, the variables which cause profits to be what they are. An issue related to the cause of profits is the moral justification of the claimant to profits. This issue though prominent in the classical and Marxian economics, disappears in neoclassical economics with the triumph of the marginal productivity theory of distribution. But the issue reappeared during the controversy surrounding capital theory in the 1960s.

Before we go into these issues, we have to ask whether profit is a pure category in economics since a theory presumes the existence, in the abstract at least, of a definable category to which the theory is addressed. With profits, there has been a frequent problem of conflating it with interest and rent.

One class of theories have taken profits as synonymous with interest. Theories have failed to distinguish consistently between profits and interest as categories of non-labour income. Thus, theories for the existence of positive rates of interest are often thought of as theories of profit rate. The abstinence theory of profits, attributed to Nassau Senior, is as much a theory of interest as of profits. In early classical writing, in Adam Smith for example, profits and interest do not appear as separate income categories. Wages, rent and profits are the three divisions of total income. When the interest rate does make its separate appearance, it is as the limit to which the rate of profit can fall. The ambiguous relationship between the two reappears in the marginal productivity theory where the return to capital (real rental on capital) is not sufficiently distinguished from interest to be a separate category of income.

There are also other conventions which lead to a loose definition of profits. Profits are sometimes used as synonymous with all non-labour income, subsuming rent and interest incomes. This is very much a Marxian tradition which has come down into modern-day Cambridge growth theory via Kalecki's theory of distribution. Alternatively, Marshallian theories subsume profits under a general category of quasi-rent. This leads to the further distinctions made between normal and supernormal profits or pure profits.

Neoclassical Theory

In neoclassical theory, the competitive firm (the entrepreneur) maximizes the quantity of profits to decide the level of output and inputs. This gives the price equal marginal cost condition. In equilibrium, the *size* of profits is indeterminate as are the *rate* and *share*. In long-run competitive equilibrium of the industry (if such an equilibrium can be shown to exist) the firm has zero (actual or if there is uncertainty, expected) profits; price equals average cost and the output is produced at the lowest point of the U-shaped average cost curve. Thus, zero profit is an equilibrium condition and also a signal that output is produced under efficient conditions. This zero profit condition is often qualified by adding that this does not rule out 'normal' profits. This not only leaves normal profits indeterminate in size but could easily lead to the condition of zero profits being a tautology. In this paradigm, non-zero profits are an indication of (long-run) disequilibrium or of non-competitive conditions (for example, barriers to entry). A third alternative is that in the presence of uncertainty any observed non-zero profits may be the random deviation of actual from expected profits. The zero profit condition is best thought of not as a descriptive prediction but as a rule to check for consistency in any model that assumes competitive behaviour.

In terms of the *rate of profit* the condition for competitive equilibrium is that the rate of profit be equal in all activities (industries, sectors, and so on). Here again the rate of profit is itself indeterminate but it is the interindustrial differential in the rate which needs to be zero in equilibrium. Again in analogy with the size of profits, a persisting non-zero differential indicates either disequilibrium or imperfectly competitive elements.

Normal non-zero level of profits or persisting differential in a particular firm or industry can be reconciled with competitive equilibrium by an appeal to Marshall's doctrine of quasi-rent. Ricardo's theory of rent compels the marginal land to have zero rent but supramarginal lands to earn positive rent. Marshall extends this logic to other factor incomes with the doctrine of quasi-rent relying on restrictions on elasticity of factor supply in the particular case where differential rates of return or non-zero surplus incomes are found. Normal non-zero profits could then be a quasi-rent.

But if so, what is the factor of production whose income is profit (as quasi-rent)? This raises the contentious question of the moral and economic basis of profit as income. Is profit the return to capital or is it the remuneration of the capitalist as a manager/entrepreneur? Attempts to provide a justification for profits as income invoke the abstinence doctrine, or the 'residual claimant' argument in the 19th century. But abstinence could provide a theory of reward for savings, that is, a theory of the interest rate but not for a reward for capital unless all savers are also capitalists. The residual claimant theory, that is, profits are what is left over after every other input has been paid its due, is hardly a theory. It needs to be supplemented by a theory of how all other inputs are rewarded, that is, how the residual is determined.

It was the achievement of J.B. Clark's marginal productivity theory of distribution to provide such a link by an attempt to integrate production and distribution via the marginal principle. Clark's theory claimed to explain distribution at each of the three levels of aggregation. The equating of the marginal value product of a factor to its unit price is a principle which treats labour and non-labour inputs symmetrically. In this theory, capital appears as equipment and its marginal revenue product is equated to its unit price. If one could further identify the profit per unit of capital with the unit price of capital, then the return to the capitalist (profits) is the reward of the productivity of capital (the rental for the factor). This brings profits in an analogous relation with rent. The capitalist is one who owns the factor of production capital. It is the structure of property rights as well as the productivity of capital which combine to make the owner of capital the recipient of its fruits.

In equilibrium, the price of the capital equipment will be the present value of its future net income stream. This is the contribution of Irving Fisher. Under certain restrictive assumptions – known future income stream, constant relative price of capital and constant and known discount factor – the real rental of capital is equal to the sum of the rate of discount (the rate of interest) and the rate of depreciation of the capital good. The presence of the rate of depreciation in the formula is required only in the case of a physical durable good. The rate of depreciation may of course not be constant but variable; worse, it could be endogenous and dependent on the forces determining the rental on capital (the rate of investment, for example). If, on the other hand, one were to take an infinitely lived capital good, that is, zero depreciation rate, then the real rate of return (the rate of profit) 'degenerates' to the rate of interest. In such a case an infinitely lived capital good becomes like a financial asset – a consol.

This illustrates the difficulty of separating profits from rent or interest. The problem here is that capital can be a physical good (a machine/a building), a financial asset (bond/consol) or an abstract attribute (human capital/skills). The marginal productivity notion relates to capital only when it is a material input to production. This is why the specification of the production function used to explain the rate of profit as arising from the marginal productivity of capital as factor becomes a contentious issue. Once capital is a physical good, its durability and heterogeneity entail an assumption of malleability if we wish to add up the disparate units of capital to arrive at an aggregate capital stock. This aggregate may be at the level of a firm, or an industry, but it is at the highest level of the economy that the problem is serious.

The need to have an aggregate measure arises from the practice of using an aggregate production function in terms of labour and capital, both assumed to be homogeneous aggregates. The use of the aggregate production function had begun in the 1920s and 1930s where on the one hand, Paul Douglas and Charles Cobb had used their well-known function to explain the constancy of the share of labour in total income (Cobb and Douglas 1928; Douglas 1948). Theorists such as Hicks, Harrod and Joan Robinson had also used the device of an aggregate production function to define various measures of technical progress which would reconcile the stylized facts of the constancy of the share labour, with a rising real wage and a trendless rate of profit. Thus it was that technical progress measures such as Hicks-neutral, Harrod-neutral, and so on, were proposed as simple constructions to explain these stylized facts. The theorists' work did not require the Cobb–Douglas form as such but the latter proved a convenient way for expressing these forms explicitly in econometric work in the 1950s when economic growth and technical progress engaged the attention of economists. It was Solow's work both in proving the stability of growth equilibrium and in measuring the contribution of technical progress to economic growth which generated the veneration of the Cobb–Douglas production function (Solow 1956, 1957).

This use of the aggregate production function in neoclassical growth theory in the 1950s accomplished two things. It could link the rate of growth of the economy to the rate of profit; in some cases the two could be equal. At the same time, the Cobb–Douglas form could be used to link the rate of profit to the marginal productivity theory. Thus, a microeconomic firm theoretic proposition – the equality of marginal productivity and factor prices – could be exploited to explain macroeconomic magnitudes of factor shares and growth.

This brilliant device linking the marginal productivity theory of distribution with a macroeconomic growth theory soon ran into difficulties and unravelled itself. The capital theory (or Cambridge–Cambridge) controversy is dealt with elsewhere in this dictionary. For the purpose of this essay, it suffices to note that while the factor prices could be taken as given by the firm, they become endo-genous at the macroeconomic level. The return to capital – the real rental or the profit rate – depends on the rate of investment. To explain the latter in terms of demand for capital in terms of factor–price substitution involves circular reasoning. Further, the existence of an aggregate capital stock presumes the prior existence of equilibrium prices for the heterogeneous capital goods comprising such an aggregate. The aggregate then cannot be used to 'explain' equilibrium profit rate. There seemed to be some insuperable logical problems in the notion of an aggregate capital stock. As an explanation of the profit rate, the macroeconomic theory of distribution proved to be a cul-de-sac.

An alternative treatment of the return to capital would be to abstract from such complications and treat capital as a commodity like any other. Using Debreu's very general definition of a commodity, new capital and old capital become different commodities, heterogeneous capital goods retain their differences and do not need to be aggregated. The task of the theory is then to show that a set of non-negative prices will clear markets for all the commodities. Given the facts of time and uncertainty we create dummy markets for contingent commodities. The price of the commodity 'capital' can then be determined for each time period and each state of nature.

But while the price of capital good can be determined by this method, it has no further link with profits. Profits in the Arrow–Debreu equilibrium are zero. Owners of capital goods will receive the price as their reward but this is not profit. Nor need the price of capital good have any specific connection with the rate of interest; as an intertemporal price the interest rate links the price of any commodity to that of its substitute available at a different date.

The Arrow–Debreu theory invokes a very stylized sort of uncertainty. Going back to a distinction made by Frank Knight between risk and uncertainty, it is risk rather than uncertainty which is involved in the Arrow–Debreu notion of contingent commodities. It assumes that states of nature can be described fully and the probabilities of various outcomes under different states of nature calculated in advance of determining the excess demand functions (or correspondences) for such contingent commodities. Such risk being previsible and insurable against, will yield only such return which cannot be arbitraged away. There can be no pure profit in such an equilibrium.

Classical Theories of Profit (Smith and Ricardo)

In classical economics profits are important as a quantity. Together with rent, they constitute the economic surplus, wages being the *faux frais* of production. The distribution of the surplus as between the owners of stock (capitalists) and the landlords becomes a central issue. This is because the two classes were presumed to have different propensities for productive consumption (accumulation). The division of surplus then had growth consequences.

The Physiocrats regarded land (nature) as the only source of surplus. Profits were only a recycled part of the *produit net*. Adam Smith could be read as regarding a surplus producing agriculture as a necessary but not the sole source of surplus, since productive labour was another source. The productivity of both land and labour depended on constant improvements in the manner of their utilization – agricultural practices and division of labour. Thus behind the 'factors' was the incessant propensity for progress – technical progress in the narrow sense as well as general innovations and improvements in practices.

The size of profits is ambiguous in Smith as there are no clear rules about the division of the surplus between rent and profit. The rate of profit was expected by Smith to fall. The reason for this was not diminishing returns in agriculture, as it was to be for Ricardo. The division of labour and effects of trade could counter the limits imposed by the soil. It was however that the stock of profitable investment opportunities was limited within a country. Smith could see that the fabulous profits made by trading companies were shrinking as merchants began to compete. Profits on industrial activities were on the moderate side. Thus the falling rate of profit hypothesis emerges to recognize an empirical fact though Smith had no reason for believing that technical progress could not stave off the tendency indefinitely.

It is David Ricardo who relates profits and rent antagonistically and has a clear view of the rate of interest as defining the lower limit of the rate of profit. In Ricardo, there is the first rigorous attempt to define the rate of profit, as a pure number which will be free of problems of valuation. By defining a one-good economy, corn, inputs and outputs can be measured in the same commodity. This is done by defining the wage in terms of corn and any equipment as product of labour. Given these two assumptions, the profit rate is defined as the ratio of net surplus – output of corn less inputs defined in terms of corn – to the capital advanced also measured by the inputs defined in terms of corn. But given the conflict in the economy between rent and profits, how was rent to be accounted for?

Ricardo defined the marginal land as zero rent yielding land. Thus the pure rate of profit could be defined on the marginal land free of complications introduced by rent. If we then add that profit rates equalize everywhere, the rent yielding land does so by achieving a superior output–input ratio compared with the marginal land. The difference is rent. Rent is an unearned income accruing to the landlords as a result of the progress of accumulation and the growth of population.

Ricardo integrates the conflict between rent and profit with a theory of growth which gives an explanation of the falling rate of profit. The size of profit as shared out with rent, determines the rate of accumulation since capitalists have a high propensity to accumulate. As accumulation proceeds, there are diminishing returns to land as well as labour. If the real wage were to be taken as constant, then the surplus above wage shrinks due to diminishing productivity of labour. But at the same time, rents rise. Hence profit is squeezed out. The rate of profit falls with accumulation.

But on the zero rent land, profits are antagonistic to wages. Thus any tendency for wages to go up will reduce the rate of profit. The determinants of real wages were accumulation (demand for labour) and growth of population (supply of labour). The Malthusian mechanism to keep real wages constant worked at least in the long run. But even in the short run, although theoretically real wages could rise with the force of accumulation, the empirical facts of population growth since the 1780s and the potentially high participation rate of men, women and children, meant that for the period in which he was writing, Ricardo could easily assume rapid adjustment of the supply of labour to a rise in the real wage rate.

New Classical Theories of Profit (Sraffa, von Neumann)

There are at least two directions in which Sraffa (1960) generalizes Ricardo. First he drops the single-good assumption but defines a standard commodity in terms of which the rate of profit is invariant to relative price changes. He also drops the assumption of a constant subsistence wage. By taking wages and profits as constituting surplus, that is, wages no longer being merely the costs of production but part of value added, the conflict between the rate of profit and the share of wages in total surplus is made explicit. The wage–profit frontier derived from the structure of the input–output information illustrates this conflict.

If we were to take Ricardo's choice of corn as the single good as a substantive and not merely a methodological device, one must attribute a large role to agriculture as the source of surplus. Ricardo is less explicit about the role of technical progress either in generating surplus or in staving off the effects of diminishing returns. By dropping land from the general model altogether as an essential input, Sraffa implicitly takes the technical conditions of production (the matrix of the input coefficients) as guaranteeing that a surplus exists. He does not however pursue the question of the disposal of profits, that is, accumulation.

In a parallel but independent work, von Neumann put forward a linear model of the economy with joint production, in which the timing of input and output was articulated. In this system, there are several processes available for producing a commodity but only those are chosen which at least yield a certain rate of surplus (say g) of the output above the inputs. Labour is one of the inputs. The converse of this proposition is that given the input and output prices, only those activities are chosen which yield the minimum rate of profit (p). If the linear system of production coefficients is indecomposable, from the duality of prices and quantities, it follows that $g = p$. If we now regard g as the rate of growth between this year's inputs and next year's output, we have in von Neumann's model, an equality between the rates of growth and profit. All profits need to be accumulated and wages have to be at subsistence level. The classical lineaments of von Neumann's model are thus clear (von Neumann 1945–1946).

Von Neumann's system leaves the size and share of profits indeterminate while making its rate determinate. It is his device of defining the production process as jointly producing final output and one year older capital goods, that is, the joint production technology that has proved a fertile innovation. In one way this device avoids the problem of heterogeneity as well as of durability of capital. Each item of capital equipment can be defined as a separate commodity and can be given a one period length of life after which it becomes another commodity, albeit a one year older version of itself. The problems of measurement of capital which plague the neoclassical aggregate production function are thus avoided. Also a strict separation is made between price of capital goods and profits. Finally the source of profits is seen as the technology which permits growth.

Technological progress is not endogenized in neoclassical economics, nor in classical economics except in a loose sense in the grand design of Adam Smith. Von Neumann as well as Sraffa leaves the question well alone. It is with Karl Marx that an ambitious model is attempted of a fully endogenous model of growth, accumulation and technological progress.

Profits as Exploitation (Marx)

Marx makes an explanation of the source of profits a central part of his theory. There are two ways of understanding this central role of profit in Marx. The *measure* of economic surplus being labour in the classical theory, an immediate question arose among certain socialist followers of Ricardo as to whether labour was also the *source* of all surplus and hence should be its *sole recipient*. While this is not the case in Smith or Ricardo, for Marx labour becomes a measure of value, and the source of surplus value. Surplus value takes the money form of profits via the mediation of prices, which are formed on the basis of values.

A second motivation of Marx's theory could be seen as extending Ricardo's critique of rent as an unearned income to the category of profits. The scarcity of fertile land is not a natural but a social phenomenon in Ricardo, caused by the progress of accumulation and population. Was the scarcity of capital and the fact that it commanded a surplus equally social? Marx asked.

Marx's theory of profits is that profits are the money form of surplus value produced by labour in the production process but appropriated by the owners of means of production. The capitalist advances capital to buy labour and means of production. But what he buys is labour power, the capacity for work. This is because the labour contract is a voluntary, *ex ante* agreement on the part of the labourer to work a fixed period of time – the length of the working day in return for a wage. The wage is the *exchange value* of the commodity labour power. The *use value* of the labour power is whatever productivity the capitalist can extract from the worker during the working day. There is a gap between the use value and exchange value of labour power but this gap cannot be seized by the sellers of labour power, but only by the buyers. This is because the buyers of labour power, the capitalists, enjoy a class monopoly of ownership of the means of production. Without finding a buyer for the labour power, the labourer cannot reproduce himself, that is, he cannot survive for any length of time with his working capacity intact. Thus, there is an asymmetry in the positions of the workers and the capitalists, as a result of a historical process that has deprived workers of direct access to means of production.

The gap between use value and exchange value present in the case of labour power is not present in the case of capital goods. These are bought and sold by capitalists and hence there is no scope for unrealized gaps to exist for any length of time without being captured by the seller. This is why Marx called the flow of input services from capital goods *constant* (c) capital, i.e. capital which had the same value in the beginning as at the end of the production process. For labour, the flow of input services – the use value realized – was made up of necessary labour measured by the exchange value, i.e. wage and surplus labour which accrued to the capitalist. This is why labour was *variable* (v) capital, it changed value between the time it was bought/paid for and the end of the production process.

The rate of profit, for Marx, was then the ratio of surplus value (s) to the sum of constant and variable capitals $(c + v)$; $[\rho = s/(c + v)]$. All the three terms of the ratio are measured in labour time and are commensurate. Thus, Marx also attempted to arrive at a measure of the rate of profit which would be invariant to relative price changes. But, unlike Ricardo, he did not assume a single good economy. What is needed is that the wage rate can be converted into labour values. The same has to be done about flow of services emanating from the means of production. This requires that the production technology be of a form which makes such conversion of goods into their labour values problem free.

The source of profits in Marx is the exploitation of labour by the capitalists, although it is subsumed that the technology is such as to yield a surplus, that is, a gap between the use value and exchange value of labour power. Positive surplus value is seen as a necessary and sufficient condition for profits to be positive (Okishio 1963; Morishima 1973). Profits are accumulated, and put into ever improving technology as a result of the competition among capitalists. It is in the constant search for higher profits, partly immanent and partly as a response to factors threatening the rate of profit, that the incessant improvement in technology takes place. Capitalists have to find better ways of increasing profits, by any means which can increase the gap between the exchange value and use value of labour power. They may do this by improving working methods (Taylorism), by extending hours of work without increasing the wage (absolute exploitation) or by investing in improved machinery (relative exploitation).

Marx in common with the classical economists also has a theory of the falling rate of profit. The progress of accumulation raises the wage rate and hence with a given technology lowers the ratio of surplus value to variable capital, that is, lowers the rate of surplus value (s/v). In order to stave off this danger, capitalists are compelled to use labour-replacing technology. This raises the organic composition of capital, that is, the share of constant capital in total capital $[c/(c + v)]$. The rate of profit varies directly with the rate of exploitation and inversely with the organic composition of capital. In the pure theoretical model, Marx reasoned that the balance would be such that the rate of profit would fall. He also discussed a number of countervailing tendencies such as the growth of monopolies in particular and a high degree of concentration in the industrial structure which may arrest this tendency of the rate of profit to fall.

Accumulation in Marx proceeds incessantly but not at a constant or equilibrium rate. The search for higher profits drives accumulation and accumulation in turn increases surplus value by being embodied in better techniques. But accumulation acts on a labour force that ultimately even exhausts its reserve army and thus wages threaten to rise. On the other hand, markets have to be found for the larger quantities of goods being produced at prices which will yield a profit, that is, surplus value has to be *realized* in the market, not just extracted from labour. The result is that the accumulation process facing these two limits results in a cyclical growth pattern. This pattern yields cycles around a declining rate of profit.

Marx's attempt to obtain a pure (relative price invariant) measure of the rate of profit has been criticized mainly due to the problem of evaluating the different types of skilled labour used in production. Marx's theory requires that all types of labour be reducible to homogeneous labour. Such reduction cannot be made without a measure of relative value productivity of different types of labour, independently of their market rates of remuneration. Despite much ingenious work, this problem has proved intractable. Another problem arises from the durability of capital. In a joint production formulation, it can be shown that positive surplus value is neither necessary nor sufficient for positive profits. The theoretical formulation has to be amended to rule out non-convexities which lead to the curiosum that negative surplus value can lead to positive profits (Steedman 1976; Morishima 1974). A third problem arises from the fact that an example of accumulation, with balanced growth and a constant rate of profit, was provided by Marx himself in his Scheme for Expanded Reproduction contradicting the necessity of cyclical accumulation or of a falling rate of profit.

Profit as Disequilibrium (Schumpeter, Keynes)

Another ambitious attempt to combine profits and growth was made by Schumpeter. With Schumpeter, profits become a *disequilibrium* phenomenon. He advances a theory of the size of profits, especially the source of profits but none of either the rate or the share of profit. Schumpeter's theory is also the only one where there is a clear link between the monetary system that finances production and the real system that generates profits. Schumpeter's is also the only theory which makes the agency of the profit earner – the entrepreneur – an explicitly central part of the theory.

The source of profits is innovation. Innovations can comprise introduction of a new good, of a new method of production, the opening of a new market, discovery of a new source of supply of raw material, or the carrying out of a new organization of an industry. The economy is supposed to be in a state of stationary or steady-state equilibrium before the innovation occurs. An entrepreneur as a visionary innovates by launching a new product or a new technique, and so on. Such a new product may have a long gestation lag before it earns revenue and, as such, may be risky. Thus in the financing of innovations, credit plays an active role. Such credit will be excess to current goods supply and will cause inflation. Once the innovation appears on the market, the entrepreneur makes monopoly profits. The credit initially created can be liquidated out of profits but the innovation causes further ripples via backward and forward linkages as well as by attracting imitators. Innovations occurring singly or in a cluster set off a long wave, a Kondratieff cycle. In the rising phase, prices, profits and output rise. But eventually the monopoly profits are bid away by competitors and the system returns to equilibrium, with profits tending to zero.

The innovation process is discontinuous and disequilibrating. It is accompanied by a credit boom and a cyclical upturn. Innovations are unanticipated. The economy exists in cycles caused by innovations but tending towards an equilibrium of zero profits, once the innovation has spent itself. The history of capitalism was for Schumpeter made up of successive long waves caused by clusters of innovations.

Thus for Schumpeter the source of profits is the superior productivity achieved by innovations but the agent of change is the entrepreneur. Neither the conventional industrialist nor labour generates profits. Profits are by nature abnormal, disequilibrium phenomena. They do not persist but dissipate in equilibrium.

Schumpeter thus reconciles a zero profit stationary equilibrium with observed facts of profits. But while the theory is an appealing one, it has lacked sufficient analytical detail to prove either a source of further developments in profit theory or a tool for empirical investigation.

While Schumpeter put forward a dynamic theory of the disequilibrium role of profits, his model is sparse in details. Keynes in his *Treatise on Money* also treats profits as disequilibrium and insists that national income calculations exclude profits. The emergence of profits as a disequilibrium category comes from the gap between savings and investment. The famous Fundamental Equations of the *Treatise* describe a two sector model with consumption goods and investment goods. For each sector, the price level is made up of unit labour cost and a disequilibrium item. For the consumption goods sector, this item consists of the cost of production of investment goods less savings. Following a similar procedure for the other sector and aggregating over the two sectors, Keynes gets the result that

$$P = w + \pi \tag{1}$$

$$\pi = (I - S)/Y. \tag{2}$$

P is the overall price level, w earnings per unit of output (unit labour cost) and π is profit per unit of output, that is, the share of profit. This identity becomes an equation only because via the expenditure equations, profits per unit of output are derived as the gap between investment expenditure (I) and savings (s) both per unit of output (Y) as in Eq. (2). In equilibrium, Keynes expects π to be zero and w to include 'normal' profits or remuneration of non-labour inputs as well as labour. When profits are non-zero this is because of investment exceeding savings and in a Wicksellian process this gap drives profits to drive the gap wider still. This process is not sufficiently articulated due to the fact that Keynes concentrates on conditions for price stability rather than on disequilibrium dynamics.

Post-Keynesian Theories of Profit and Growth

Kalecki's theory marks a bridge between Marxian and Keynesian traditions and is the seminal contribution to what is now called the Post-Keynesian or Cambridge theory of income distribution. Paradoxically its points of contact with the *Treatise on Money* have not been sufficiently brought out. Kalecki's route was via Rosa Luxemburg's critique of Marx's Schemes of Expanded Reproduction (SER). The SER is also a two-sector model but of a growing economy. It has a two goods/two class configuration which is similar to the Fundamental Equations of the *Treatise*. While Marx's formulation of the SER make the model an equilibrium one, Luxemburg was seeking to find roots of dynamic disequilibrium within it. There are several strands which Kalecki weaves into this story.

Kalecki has a macroeconomic theory of pricing which yields a determinate share of profits in total output. He does this by exploiting the marginal revenue equals marginal cost conditions of equilibrium for the neoclassical firm. By then exploiting the simple idea that the ratio of price to marginal revenue departs from one to the extent that the price elasticity of demand is below infinity he connects price to marginal cost via the demand elasticity. Thus

$$p = mc\left(1 + \eta^{-1}\right) \tag{3}$$

where mc is the marginal cost and η is the elasticity of demand. The coefficient $(1 + \eta^{-1})$ is called the degree of monopoly. To the extent that η^{-1} departs from zero, the firm is a monopolistic one.

This is a partial equilibrium, microtheoretic derivation of the p/mc ratio and its generalization to a macro-economic level has proved to contain problems (Mitra 1954). The main problem is that if (3) is supposed to refer to a specific firm, its elasticity of demand is not a constant but a function of the firm's own and its rivals' strategies. A determinate and tractable aggregation procedure for many jointly dependent p/mc ratios is not possible. It has however been found possible and empirically fruitful to interpret pricing decision as a mark-up above average cost.

$$p = (1 + k)ac \tag{4}$$

where ac is average cost and k is the mark-up ratio. The similarity of (4) to Keynes's Fundamental Equation in (1) is striking, that is, $\pi = k/(1 + k)$. But while (1) is an identity, (4) could be thought of as an equation where the profits come from producers' price setting behaviour.

But how are these profits sustained or in Marx's terminology realized? This is where the aggregate demand relations become important. It would be through the spending behaviour of the profit receivers that profits can be sustained. This was already clear in Keynes's invocation of the widow's cruse parable whereby a Wicksellian cumulative dynamic process can sustain growing profits as long as capitalists spend (that is, dissave) while keeping up their investment expenditure. By starting with the Marxian SER, Kalecki was able to derive this as an *equilibrium* relation.

Kalecki's macroeconomic theory is best seen in terms of Kaldor's generalization. Kaldor takes the two class/two good model and integrates profits into a theory of growth and distribution. Let R be total profits ($\equiv \pi Y$) and W be the total wage bill ($= wL$). Then

$$Y = R + w \tag{5}$$

$$I = S = s_w W + s_c R. \tag{6}$$

Equation (5) is a national income identity, whereas (6) combines the Saving–investment equality with a decomposition of total savings into workers' savings ($s_w W$) and capitalists' savings ($s_c R$) with the $s_c s_w$ being saving propensities and $s_c > s_w$. From (5) and (6), we can derive

$$\begin{aligned} R/Y &= \pi \\ &= (s_c - s_w)^{-1}(I/Y) - s_w(s_c - s_w)^{-1} \end{aligned} \tag{7a}$$

and

$$\begin{aligned} R/K &= \rho \\ &= (s_c - s_w)^{-1}(I/K) - s_w(s_c - s_w)^{-1}(Y/K). \end{aligned} \tag{7b}$$

Equation (7a) gives the share of profits in terms of the investment income ratio and (7b) gives the rate of profits in terms of the rate of growth of capital stock (I/K) and the output–capital ratio (Y/K). To specialize the equation, set $s_w = 0$, that is, assume workers do not save. Then

$$\pi = {s_c}^{-1}(I/Y) \tag{8a}$$

$$\rho = {s_c}^{-1}(I/K). \tag{8b}$$

These two equations show how the profit share is determined by the capitalists' investment behaviour, i.e. capitalists determine their own profits. If we take the output–capital ratio to be a constant, then the rate of growth of income (g) is equal to the rate of growth of capital stock (I/K). Thus from (8b), we have

$$\rho = {s_c}^{-1} g. \tag{9}$$

If we now put $s_c = 1$, we have the von Neumann result reproduced in the Kalecki–Kaldor models.

The restriction that $s_w = 0$ is of course arbitrary and thus makes the result under (8b) somewhat unrealistic. Pasinetti (1962) generalized the Kaldor argument by allowing workers as well as capitalists to save and own capital. Thus total capital K could be held either by capitalists K_c or by workers K_w but since capitalists make output and investment decisions workers were assumed to have loaned K_w to capitalists. In terms of the distinction we made above, capital as productive equipment is *controlled* by capitalists but capital as a financial asset is *owned* by both workers and capitalists, and capitalists pay workers a rate of interest i on the loaned capital. Thus instead of (7a) and (7b), we get

$$\begin{aligned} R/Y &= \pi \\ &= (s_c - s_w)^{-1}\Big[(I/Y - s_w) + r\Big(s_w s_c (I/K)^{-1} \\ &\quad - s_w (Y/K)^{-1}\Big)\Big] \end{aligned} \tag{9a}$$

$$\begin{aligned} R/K = \rho = (s_c - s_w)^{-1}\Big[&(I/K - s_w(Y/K)) \\ &+ r\Big(s_w s_c (I/K)^{-1} - s_w\Big)\Big] \end{aligned} \tag{9b}$$

If we now put $r = \rho$, (9a) and (9b) degenerate to

$$\pi = {s_c}^{-1}(I/K) \tag{10a}$$

$$\rho = {s_c}^{-1}(I/K). \tag{10b}$$

The only condition needed for this result is that $(I - s_w Y) = 0$. But while (10b) is similar to (8b), now it is independent of whether s_w is zero or not.

The similarity of the Kalecki–Kaldor result to the von Neumann result, as we noted above, is striking. The Pasinetti result seems to reinforce it. It is a one-good model and hence problems of relative price or aggregation or measurement of capital which plague other theories are completely avoided here. It is also not clear as to whether causality proceeds from growth to profits or profits to growth. There is an implicit assumption that the economy must have adequate resources and technology to generate surplus but the source of the surplus is not clear. There is no specification of the production conditions and a neoclassical aggregate production function is deliberately avoided.

The Pasinetti result has been derived by an alternative route by Samuelson and Modigliani (1966) who do use a neoclassical aggregate production function. Their purpose was to point out that the Pasinetti result was a special case of a more general result and that a dual to Pasinetti's theorem – an anti-Pasinetti theorem – could be derived from a slightly alternative formulation. All the assumptions of Pasinetti's theory are retained except that profits and wages are now derived from the marginal productivity conditions and a constant return to scale, two factor production function.

Let the production function be

$$\overline{Y} = f\left(\overline{K}\right) f' > 0, f'' > 0. \tag{11}$$

Here $\overline{Y} = Y/L, \overline{K} = K/L$, that is, output per worker and capital per worker. By the standard rules of marginal productivity theory we have that wage and rate of profit are determined as

$$\rho = f'\left(\overline{K}\right) \tag{12a}$$

$$w = f - \overline{K}f'\left(\overline{K}\right). \tag{12b}$$

In the production function, there is no distinction as to who owns the total capital stock – capitalists or workers. The savings augment the amount of capital owned by workers and capitalists,

$$S_c = \dot{K}_c = S_c\, f'(K)\, K_c \tag{13a}$$

$$S_c = \dot{K}_w = s_w[Y - f'(K)K_c]. \tag{13b}$$

The equilibrium condition in the Samuelson–Modigliani model is that the relative rates of growth of capital stock owned by workers and capitalists be the same, i.e. constancy of shares in productive wealth. This is not an obvious condition for equilibrium but it does have the dramatic consequence that in such an equilibrium (of balanced growth of K_c and K_w), the rate of profit is independent of the saving propensity of the worker. If n is the constant growth rate of the capital stock, we get from the above after some manipulation

$$\dot{K}_w = s_w[f(K) - K_c f'(K)] - nK_w \tag{14a}$$

$$\dot{K}_c = [s_c f'(K) - n]K_c. \tag{14b}$$

In steady state $\dot{}K_c = \dot{}K_w = 0$, so (14b) gives

$$\rho = f'(K) = n/s_c. \tag{15}$$

Thus, the Pasinetti result can be derived from a neoclassical logic. This should not be too surprising though much was made of this paradox at the time (Pasinetti 1966; Kaldor 1966). Neither a condition such as $\rho = r$ (what we have called the degeneracy result) nor that K_c/K_w is a constant tell us very much about the mechanisms by which an economy can arrive at such results. Our economic world is a world of many heterogeneous goods – capital as well as labour, of uncertainty, of financial constraints, of the persistent possibility of technical progress, of mergers and takeovers. There is in these models no decision making agency and time is eliminated in any meaningful sense since no *ex ante* versus *ex post* distinctions can be made. The Kalecki–Kaldor–Samuelson–Modigliani propositions are simple parables of pedagogic value no doubt but they do not tell us much about the origins or the role of profits in a modern economy.

Behavioural Theories of Profit

We can move in the final section of our essay to theories where the behavioural context is much more explicit. The neoclassical firm is a black box, where all the allocative rules could be followed by a computer which can be programmed to equate a derivative to a price. But the modern economy consists of corporations which operate in a world of financial markets and profits and are in this world both a signal of managerial performance and a facilitator for future expansion. It is this cluster of theories which supply the missing dimension in the theories hitherto surveyed.

Entrepreneurial Theories of Profit

In this cluster of theories, the level of aggregation is the firm and the agency of decision making is identified as the entrepreneur. It is assumed that the firm operates in a noncompetitive but stable environment where the entrepreneur has to form expectations about economic variables within his control as well as rivals' strategies. The other set of variables is the macroeconomic one where entrepreneurs may hold similar short run expectations. The level of profit sustained by the firm will depend in the short run on macroeconomic factors, common to all firms but in the long run on the decisions to invest so as to keep rivals at bay (Keirstead 1953).

Knight's definition of uncertainty and Keynes's view of the difficulty of rational calculus in forming long-run expectation have been synthesized by Shackle in a series of works (Shackle 1954, 1969, 1970) which purport to relate investment and *ex ante* profit in a decision theoretic framework. Shackle's decision theory is not however that of von Neumann–Morgenstern. He puts forward the notion of potential surprise and a surprise function relating the 'size' of potential surprise to the profit or loss attached to the project. Along with the surprise functions there is an ascendancy function which relates the entrepreneur's engagement with a project given the size of gain or loss anticipated.

Shackle then describes an optimizing exercise on the part of the entrepreneur, yielding optimal potential surprise. The size of gain or loss attached to the optimal potential surprise is then called primary focus gain or loss. The zero surprise (that is, certainty) equivalent of this primary gain is what Shackle calls profits. Profits are thus an *ex ante* certainty equivalent measure of the likely gain of the optimally chosen investment project. This definition unlike all previous ones makes the subjective and *ex ante* concept of profits clear. The problems with Shackle's theory are that it is determinedly resistant to aggregation even over individuals and while emphasizing the difficulty of using the calculus of subjective probability it assumes that the surprise function and the ascendancy functions are continuous and differentiable enough to define a unique maximum. There is obviously a contradiction here.

Mention should be made in this respect of Lamberton's work (Lamberton 1965). Lamberton prefers a satisficing to a maximizing approach and explicitly introduces the entrepreneur's income–leisure (inactivity) trade-off as determining the choice of investment projects and the associated profit outcome. In later work, amberton rationalizes suboptimizing behaviour in terms of the costs of information gathering and processing. Entrepreneurial activity to maintain or enhance profits is triggered off subject to threshold effects.

A Corporate Theory of Profit

Authors such as Keirstead and Shackle view the firm as a single-owner entity. Modern firms are not by and large individually owned. In terms of share of output, employment or sales, it is the corporation with team management and hierarchical command structures which is the dominant mode of firm organization. It is not surprising therefore that one class of theories deals with profits in the context of corporate or managerial behaviour.

While in general Galbraith can be said to have made economists aware of the corporate form, corporate theories could be said to have benefited from the contribution of the behavioural theory of firm of Simon, Cyert and March, the hypothesis attributed to Baumol that firms maximized sales rather than profits and the notion due to Marris among others that growth of corporate size was the aim of managers of corporations (Baumol 1967; Cyert and March 1963; Galbraith 1967; Marris 1964; Simon 1957). These various developments in industrial economics brought forth a view that managerial behaviour was growth oriented, that operating in monopolistic competitive environments managers could choose profits and growth combination subject to the constraint of external sources of finance.

Wood (1975) presents the most complete theory in this respect. There are no details on the production side in his theory of profits. Profits arise from the need of the corporation to grow. The choice variables are the retention ratio, that is, the proportion of investment financed from external sources and the amount of liquid financial assets desired by the firm to be held as some relation of gross investment expenditure.

The firm is assumed to be facing a convex opportunity frontier between the profit share (profit margin as Wood calls it) and the growth rate of its sales revenue. Its desired (or target) values of the retention ratio (γ) the external finance ratio (λ) and the liquid financial holdings ratio (δ) give a simple relation between profit share (π) and the growth rate g. We have

$$I + \delta I \leq \gamma R + \lambda I. \tag{16}$$

In Eq. (16), the left-hand side represents the uses of funds and the right-hand side the sources of funds. A slight rearrangement gives us

$$R/Y = \pi \geq \gamma^{-1}(1 + \delta - \lambda)\sigma g \tag{17}$$

where $g = \Delta Y/Y$, the growth rate of sales and $\sigma = I/\Delta Y$ is investment to increase in output (or the incremental capital output) ratio. Now if one could accept that the coefficients δ, λ and γ are fixed parameters and σ is also constant, then π and g are linearly related. The firm will start with a desired value of g and seek the appropriate Π and by an iterative process arrive at a Π, g combination which satisfies (17) as an equality rather than an inequity. This Π, g combination will also be on the boundary of the opportunity frontier.

Wood's theory treats the corporation in isolation and with some control over its pricing and revenue situation. The presence of competing monopolistic firms is ignored here as in Kalecki's theory. But his theory does bring out the interrelationship between profits and the financing of investment. Note also that in (16) all the variable are in nominal terms. If the parameters δ, λ and γ were identical across firms, we would aggregate (16) across firms. Since Y is sales revenue in nominal terms Eq. (17) would make sense at an aggregate level though it would be harder to swallow that σ would be a constant.

Conclusion

A satisfactory theory of profits is still elusive. For neoclassical economics, profit is a non problem and the only problem is to assign any observed net income above costs to the category of interest, quasi-rent or managerial wages for risk bearing. But problems persist for other theories as well. The classical theory neglects uncertainty and is vague about the microfoundations of profits. If von Neumann and Sraffa detail the technical conditions of production, neither uncertainty nor demand considerations figure prominently in their theories. Most theories, with the exception of the corporate theories and Schumpeter's, neglect the financial aspects of business operations. The firm is a black box in neoclassical theory whereas for Knight, Keynes and Shackle the individual decision maker's expectations are crucial to business behaviour. To allow for persistent positive profits in a dynamic equilibrium microeconomic model with subjective uncertainty and expectations with the possibility of technical substitution and technical change in production remains a challenge. If such a theory could be cast in terms to allow aggregation over firms to the level of the economy, the puzzle of profits would be solved.

See Also

- ▶ Worker Participation and Profit Sharing

Bibliography

Baumol, W.J. 1967. *Business behaviour, value and growth*. 2nd ed. New York: Macmillan.

Carter, C.F., and J.L. Ford, eds. 1972. *Expectations and uncertainty in economics*. Oxford: Blackwell.

Cobb, C., and P. Douglas. 1928. A theory of production. *American Economic Review* 18: 139–165.

Cyert, R., and J.G. March. 1963. *A behavioural theory of the firm*. Englewood Cliffs: Prentice-Hall.

Douglas, P. 1948. Are there laws of production? *American Economic Review* 38: 1–41.

Galbraith, J.K. 1967. *The new industrial state*. Boston: Houghton Mifflin.

Kaldor, N. 1955. Alternative theories of distribution. *Review of Economic Studies* 23 (2): 83–100.

Kaldor, N. 1966. Marginal productivity and the macroeconomic theories of distribution. *Review of Economic Studies* 33: 309–320.

Kalecki, M. 1939. *Essays in the theory of economic fluctuations*. London: Allen & Unwin.

Keirstead, B.S. 1953. *An essay in the theory of profits and income distribution*. Oxford: Blackwell.

Keynes, J.M. 1930. *A treatise on money.* Vol. 1. London: Macmillan.

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

Lamberton, D.M. 1965. *The theory of profit*. Oxford: Blackwell.

Lamberton, D.M. 1972. Information and profit. In Carter and Ford (1972).

Marris, R. 1964. *The economic theory of managerial capitalism*. London: Macmillan.

Meek, R.L. 1954. Adam Smith and the classical concept of profit. *Scottish Journal of Political Economy* 1 (2): 138–153.

Mitra, A. 1954. *The share of wages in national income*. Rotterdam: Netherlands Central Planning Bureau, limited circulation edition. Calcutta: Oxford University Press, 1980.

Morishima, M. 1973. *Marx's economics: A dual theory of value and growth*. Cambridge: Cambridge University Press.

Morishima, M. 1974. Marx in the light of modern economic theory. *Econometrica* 42: 611–632.

Okishio, N. 1963. A mathematical note on Marxian theory. *Weltwirtschaftliches Archiv* 91 (2): 287–299.

Pasinetti, L. 1962. Rate of profit and income distribution in relation to the rate of economic growth. *Review of Economic Studies* 29: 267–279.

Pasinetti, L. 1966. New results in an old framework. *Review of Economic Studies* 33: 303–306.

Samuelson, P.A., and F. Modigliani. 1966. The Pasinetti paradox in neoclassical and more general models. *Review of Economic Studies* 33: 269–302.

Schumpeter, J.A. 1912. *Theorie der Wirtschaftlichen Entwicklung, Eine Unterschung uber Unternehmergewinn, Kapital, Kredit, Zins und den Konjukturzyklus*. Munich and Leipzig: Duncker & Humblot. Trans. as *Theory of economic development*. Cambridge, MA: Harvard University Press. 1934.

Shackle, G.L.S. 1954. Professor Keirstead's theory of profits. *Economic Journal* 64: 116–123.

Shackle, G.L.C. 1969. *Decision, order and time in human affairs*. 2nd ed. Cambridge: Cambridge University Press.

Shackle, G.L.C. 1970. *Expectation, enterprise and profit*. London: Allen & Unwin.

Simon, H.A. 1957. *Models of man: Social and rational*. New York: Wiley.

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

Solow, R.M. 1957. Technical change and the aggregate production function. *The Review of Economics and Statistics* 39: 312–330.

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

Steedman, I. 1976. *Marx after Sraffa*. London: New Left Books.

von Neumann, J. 1945–6. A model of general equilibrium. *Review of Economic Studies* 13, 1–9.

Wood, A.J.B. 1975. *A theory of profits*. Cambridge: Cambridge University Press.

Progressive and Regressive Taxation

William Vickrey and Efe A. Ok

Abstract

Progressive (resp. regressive) taxation refers to a taxation scheme in which the amount of tax paid as a proportion of the tax base rises (resp. declines) with that base. While progressive taxation has been justified in terms of the 'principle of equal sacrifice' and mitigating the inequality of market outcomes, no general political theory of income taxation provides theoretical support for the observed prevalence of progressive taxation schemes. Nor has the theory of optimal income (and consumption) taxation shed any light on the nature of progressive taxation, either normatively or positively.

Keywords

End-point theorem; Equal sacrifice principle; Equity–efficiency trade-off; Income mobility; Jakobsson–Fellman Theorem; Lorenz curve; Mixed strategy equilibrium; Optimal taxation;

Progressive and regressive taxation; Redistribution of income and wealth; Reynolds–Smolensky progressivity index; Tax base; Tax burden; Tax incidence

JEL Classifications
H2

Progressive (resp. regressive) taxation refers to a taxation scheme (applied to a monetary base such as income, consumption, wealth and so on) in which the amount of tax paid as a proportion of the tax base rises (resp. declines) with that base. As such, characterizing taxes as progressive or regressive according to the relative degree to which they impose burdens on the wealthy and on the poor seems at first blush a fairly straightforward matter. Unfortunately, when one wishes to understand the nature of *effective* progressivity of a taxation scheme, and its redistributive properties thereof, and especially to compare the relative degree of progressivity or regressivity of alternative taxes, a number of problems arise. Before formalizing the notion of progressivity, we thus have a quick look at these issues.

Tax Base

An important and often overlooked question is that of the base to which the tax burden is to be related. Usually the base is taken to be income, which in practice, whenever an attempt is made to produce actual figures, means some version of annual monetary income (which perforce includes salaries, interest and dividends, but which may or may not include capital gains). An alternative basis for evaluating progression would be consumption. This is not often used, and is subject to the same deficiencies as income as a result of omissions of non-monetary items such as imputed incomes from consumer durables (like owned homes).

Tax Burden

There are various difficulties in determining the actual tax burden levied on a taxpayer. First, application of tax rates to incomes that fluctuate (due to, say, the stage of one's life cycle) makes this difficult to measure. As advocated by Vickrey (1947), it appears that some form of income averaging is needed (to serve as a proxy for the expected permanent incomes) but this has been found to be too complicated to administer in practice. Second, it is not obvious how to incorporate family size in the computation of tax burdens. Simply taking the aggregate would misclassify the larger and smaller units relative to their level of welfare or ability to pay, while taking a per capita measure overstates the importance of children relative to their needs. (See Pechman 1987, pp. 78–133, for a careful discussion of these issues and other structural problems with tax assessment in general.) Third, the market mechanism often allows the tax burden to be passed from the taxpayer to other units in the economy, with the eventual consequence that the burden of taxes is not necessarily borne by those upon whom they are levied. In particular, the imposition of taxes on income or sales changes the budget sets of individuals, thereby altering equilibrium prices in the economy. The issue becomes particularly pressing in the case of corporate income tax, as tax incidence in this case varies widely according to fiscal, monetary or activity level changes that are associated with an alteration in the tax. For instance, a tax imposed on firms for hiring labour is more than likely to be 'shifted' to workers through lower wages and to consumers through higher prices, upon the adjustment of employment decisions on the part of the firms. In general, then, the ultimate distribution of the tax burden – the so-called *economic incidence* – is different from *statutory incidence*, that is, the initial distribution of tax liabilities. Unfortunately, it is a highly non-trivial matter, both empirically and theoretically, to determine precisely who bears the tax burden, and ultimately to what extent. It is thus not uncommon in practice to encounter discussions on tax progressivity that ignore tax incidence problems and concentrate instead simply on statutory incidence. (See Musgrave and Musgrave 1989, chs. 12 and 13, for an introductory account of tax incidence analysis, and Kotlikoff and Summers 1989; Atkinson 1994, for advanced treatments.)

It is worth noting that the issues concerning the base and burden of taxes are integral to any sort of fiscal analysis, and are not particular to the analysis of tax progressivity. To flesh out the elements of the latter, therefore, we shall abstract from these difficulties in what follows and assume that a notion of monetary outcome, which we shall simply refer to as *income*, is determined as the arbiter of ability to pay. Moreover, for the most part we shall work under the (uncomfortable) supposition that the amount of tax charged on a given level of income corresponds to the actual tax burden of the taxpayer with that amount of income. At the very least, this will allow us to focus properly on certain facets of the *theory* of progressive taxation.

Progressive (Regressive) Tax Functions

Formally speaking, a *tax function* is a right-differentiable and strictly increasing map $T : \mathbb{R}_+ \to \mathbb{R}_+$ such that $T(0) = 0$ and $T(x) < x$ for all $x \in \mathbb{R}_{++}$, and $0 < T'(x) < 1$ for all $x \in \mathbb{R}_+$. (Here we denote the right-derivative of T by T'.) This formulation maintains that (i) zero income earners do not pay any taxes; (ii) if a person earns positive income, the amount of taxes imposed on her must be less than her taxable income base; (iii) higher-income earners pay a higher level of tax than lower-income earners; and (iv) taxation is non-confiscatory in that the ranking of taxpayers by pretax income and post-tax income is the same. (We rule out here negative taxation to simplify our exposition, and view T as modelling a *statutory* tax scheme.) A tax function T is said to be *progressive* if the map $x \mapsto \frac{T(x)}{x}$ is increasing on $\mathbb{R}_{++}$, that is, if the amount of income tax paid as a proportion of the tax base (say, income) rises with that base. In turn, T is *regressive* if $x \mapsto \frac{T(x)}{x}$ is decreasing on $\mathbb{R}_{++}$. Finally, T is *marginal-rate progressive* (*regressive*) if the tax rate T' is itself an increasing (decreasing) function. In practice, statutory taxes on income and spendings are always progressive – in fact, they are almost always marginal-rate progressive – while payroll and sales taxes possess a flat statutory tax rate (but economic incidence analyses frequently reveal that such taxes are, effectively, regressive).

Normative Basis for Progressivity

The most well-known equity principle – advanced originally by John Stuart Mill – that provides a normative basis for progressive taxation is the *principle of equal sacrifice.* The modern formulation of this principle demands that there be a social norm, represented by a continuous, concave and strictly increasing (social) utility function $U : \mathbb{R}_+ \to \mathbb{R}$ relative to which the income tax T imposes equal sacrifice upon all taxpayers, that is,

$$U(x) - U(x - T(x)) = \text{constant for all } x > 0$$

(see Young 1987, 1990; Ok 1995). We may now ask: does a progressive, or a marginal-rate progressive, tax necessarily satisfy the principle of equal sacrifice? Conversely, does this principle necessitate progressivity?

The answers are, unfortunately, not very clear. First, the good news: it can be shown that a marginal-rate progressive tax function surely satisfies the principle of equal sacrifice. The bad news is that mere progressivity of a tax function is not enough for it to satisfy this principle (Mitra and Ok 1997). To make matters worse, even for some non-progressive taxes T we can find a (social) utility function U that satisfy the properties above, that is, the principle of equal sacrifice need not imply, or be implied by, tax progressivity. At the very least, one needs to assume more about T and U to be able to relate these principles more closely. For instance, if we demand that U be differentiable (at least near the origin), then a piecewise linear tax function T satisfies the principle of equal sacrifice (as we formulated above) *if, and only if*, T is marginal-rate progressive (Mitra and Ok 1996). If one is prepared to accept this set-up, therefore, the principle of equal sacrifice can be thought of as characterizing marginal-rate progressivity of a (statutory) tax function, thereby necessitating its progressivity.

An additional caveat here is, of course, that this account ignores the disincentive effects of taxation. It is partial relief that Berliant and Gouveia (1993) have shown that, when the individual utility functions over income and leisure are additively separable, the link between the principle of equal sacrifice and progressivity would prevail even in the presence of such effects. Unfortunately, little is known about this matter in the (more realistic) non-separable case.

Redistributive Consequences of Progressivity

One of the traditional arguments for progressive taxation is that such schemes redress the highly inegalitarian outcomes of the market system, thereby acting as social insurance against inequality. As colourful as it may be, this argument needs to be formalized properly.

Let us first agree to model an income distribution as a continuous and increasing distribution function $F : \mathbb{R} \to [0, 1]$ with $F(0) = 0$ and $F(1) = 1$. This sort of a specification is, for instance, frequently adopted in macroeconomic models of income distribution. For any such F, we let $\mu_F := \int_0^1 x dF(x)$, which is the total income in the society. (Since incomes are distributed on $[0, 1]$, that is, we effectively concentrate on relative incomes, μ_F also corresponds to the per-capita income in this model.) In what follows we naturally assume that $\mu_F > 0$.

For any such income distribution F, the *pseudo-inverse* of F is defined as the function $F^{-1} : (0, 1) \to \mathbb{R}_+$ with

$$F^{-1}(t) := \inf\{x \geq 0 : F(x) \geq t\}, 0 < t < 1.$$

Intuitively, we may think of $F^{-1}(t)$ as the income level of the person who belongs to the poorest $100t$ per cent in the income distribution. We next define the map $L_F : [0, 1] \to \mathbb{R}$ by

$$L_F(p) := \frac{1}{\mu_F} \int_0^p F^{-1}(t)\, dt, 0 \leq p \leq 1,$$

which corresponds to the cumulative share of income held by the poorest 100 per cent of the population. The graph of the map $p \mapsto L_F(p)$ is called the *Lorenz curve* of the distribution F.

We say that the income distribution F *Lorenz dominates* another income distribution G whenever $L_F(p) \geq L_G(p)$ holds for all $p \in [0, 1]$, with strict inequality for at least one p. It is well known that this happens if, and only if, G can be obtained from F by means of finitely many mean-preserving spreads (Rothschild and Stiglitz 1970). This is one of the reasons why Lorenz dominance is generally accepted as an *unambiguous* method of making ordinal inequality comparisons. Its welfare basis is identified in the seminal works of Kolm (1969) and Atkinson (1970).

Now take any income distribution F. A tax function T applied to this distribution induces the *post-tax income distribution* F^T where $F^T(x){:} = F(x - T(x))$ for any $x \in \mathbb{R}$. A celebrated theorem of public economics – often called the *Jakobsson–Fellman theorem* – maintains that F^T Lorenz dominates F, that is, tax is inequality-reducing, if, and only if, T is progressive (see Fellman 1976; Jakobsson 1976). That is, progressive taxes, *and only progressive taxes*, possess the property of reducing the level of income inequality no matter to which pre-tax income distribution they are applied. (For variations on this theme, see Eichhorn et al. 1984, and Thon 1987.) This shows, in a nutshell, why the progressivity of a taxation scheme may be justified on the basis of desire for inequality reduction.

The Jakobsson–Fellman theorem also leads to a natural method of quantifying the redistributive effect of a tax function T that is applied to an income distribution F. To see this, let us consider the function $R_{F,T} : [0, 1] \to \mathbb{R}$ defined by

$$R_{F,T}(p) := L_{F^T}(p) - L_F(p).$$

In words, $R_{F,T}(p)$ measures the income share of the poorest $100p$ per cent in excess of what they would obtain under an equal yield flat tax. Obviously, the Jakobsson–Fellman Theorem says that $R_{F,T} \geq 0$ for *any* income distribution F if, and only if, T is progressive. (But, of course, we may have $R_{F,T} \geq 0$ for *some* F even if T is not progressive.) Now the discussion above suggests to declare a tax function T^1 to be *more redistributive than* T^2 – due to the Jakobsson–Fellman theorem, one often says T^1 *is more progressive than* T^2 – relative to the income distribution F if$R_{F,T^1} \geq R_{F,T^2}$. This is, of course, a partial ordering, and when it does not apply one may wish to resort to a compatible index that sizes up the redistributive effects of T^1 and T^2. The most widely used index to this effect is the *Reynolds–Smolensky progressivity index* defined (as a function of a tax function T) by

$$I_F^{\mathrm{RS}}(T) := 2\int_0^1 R_{F,T}(p)\,dp,$$

which is the difference between the Gini coefficients of the distributions F and F^T. Other indices are proposed in the literature to compare the progressivity of tax functions (which are based, for instance, on the departure of a tax function from the equal-yield proportional tax). For an extensive discussion of such indices, and further results on the redistributional effects of progressive taxes, we refer the reader to the excellent survey by Lambert (1999).

Political Economy of Progressive Taxation

Now that we have examined a number of normative rationales for progressivity of income taxes, let us turn to the strand of literature that has attempted to explain the prevalence of such tax schemes from the viewpoint of behavioural political economy. This literature maintains that, given that her views about income tax policy is one of the most important traits of a political candidate, it is natural to expect this prevalence to reflect (however indirectly) the majority support in the population. In fact, this way of thinking seems to suggest a straightforward explanation of the empirically observed popularity of marginal-rate progressivity, provided that one subscribes to the 'one-man one-vote' rule. Since the income distribution of a country is always globally right-skewed (in the sense that the median income is strictly smaller than the mean income for any right truncation of the income distribution), the number of poorer voters always exceeds the number of richer voters, regardless of how one defines the cut-off that separates the poor from the rich. Since poorer voters are typically the supporters of progressive policies, so the argument goes, there would then be a natural tendency for the marginal-rate progressive tax policies to be favoured by the majority. Even though the actual political processes are far more complex than the scenario in which people vote directly over policies, this argument appears to suggest a convincing reason for why progressive tax policies are so widely adopted.

While there are a few direct democracy models in the literature that provide support for this argument (cf. Romer 1975; Roberts 1977; Cukierman and Meltzer 1991; Gouveia and Oliver 1996; Marhuenda and Ortuño-Ortín 1998; Roemer 1999), these models are either confined to rather specific settings (in which a tax function is characterized by means of at most two parameters) or are not couched in a political equilibrium framework. A natural direct democracy model of voting over income taxes would be a two-party voting game in which each party (whose objective is to win the elections) proposes a tax function from an exogenously given set of admissible tax functions (that raise a given amount of revenue), and voters vote selfishly for the tax function that taxes them less. To make transparent the difficulties that pertain to the political economic approach to progressive taxation, we now describe such a voting model in precise terms.

Let F be a strictly increasing income distribution (as modelled above), and assume that the median income is strictly less than the average income according to F, that is, $m_F := F^{-1}\left(\frac{1}{2}\right) < \mu_F$. (This assumption is but a straightforward formalization of the heuristic statement that 'the number of poor people in the society is strictly less than that of rich people'.) To focus on the issue of *redistribution*, it is assumed that tax policies are designed to collect an exogenously given amount of revenue $0 < \alpha < \mu_F$, or put differently, an *admissible tax function* T is defined in the model as one with the property $\int_0^1 TdF = \alpha$. We denote the class of all such tax functions by $\mathscr{T}_{F,\alpha}$.

Consider a two-party voting model in which each party advocates an income tax policy in $\mathscr{T}_{F,\alpha}$ which is to be put in effect in case this party obtains the support of the majority. Citizens evaluate proposals from a selfish point of view. Put precisely, an individual with income x regards the tax function

T^1 as more desirable than the tax function T^2 if $T^1(x) < T^2(x)$, that is, if this individual's tax liability is lower under T^1. It is assumed that indifferent voters abstain from voting. Thus, if party 1 proposes tax policy t_1 and party 2 proposes tax policy t_2, the share of votes obtained by the first party is determined as

$$\omega\left(T^1, T^2\right) := p_F\left\{x \in [0, 1] : T^1(x) < T^2(x)\right\},$$

where p_F is the probability measure induced by F on [0, 1]. Of course, in this case the share obtained by party 2 is ω (T^2, T^1). The formal model takes the form of a two-person strategic game in which the two players are the parties both of whose action spaces equal $\mathscr{T}._{F,\alpha}$ There is a multitude of ways of modelling the objectives of the parties here. We follow Carbonell-Nicolau and Ok (2007), and presume that the goal of party i is the maximization of the net plurality defined as the difference between the vote shares obtained by the candidates. (For instance, the payoff function of party 1 is the map $\left(T^1, T^2\right) \mapsto \omega\left(T^1, T^2\right) - \omega\left(T^2, T^1\right)$ on $\left(\mathscr{T}_{F,\alpha} \times \mathscr{T}_{F,\alpha}\right)$.

While this model is one of the simplest of its kind, it readily exhibits the familiar difficulty of (infinite-dimensional) voting games: it does *not* have a Nash equilibrium (for any given F and α). Intuitively speaking, this is because, given any admissible tax T in $\mathscr{T}_{F,\alpha}$, one can always find another tax function which is below T over an interval of p_F measure greater than one half. Not all hope is lost here, however, as it can be shown that there is at least one mixed strategy equilibrium of this game. The question then becomes whether or not the support of any such equilibrium consists only of progressive taxes. Curiously, Carbonell-Nicolau and Ok (2007) show that, if F and α satisfy a certain condition, then, generically, in at least one equilibrium the probability of parties proposing a non-progressive tax function is positive. All in all, after a lot of work, and absent a good economic reason for working with a particular way of restricting the set of admissible tax functions, one is left with the feeling that the prevalence of tax progressivity that we find in all industrial democracies cannot be attributed solely to the right-skewedness of income distributions.

In passing, we should note that, in the context of *representative* democracies (à la Osborne and Slivinski 1996; Besley and Coate 1997), one is sometimes able to escape from the equilibrium existence problem discussed above. Indeed some positive results on the majority support for progressive taxation are obtained in this setting (cf. Carbonell-Nicolau and Klor 2003). Unfortunately, it is not known whether or not these results would survive the inclusion of disincentive effects of taxation into the model. Furthermore, if we add to the picture dynamic considerations of voters, things would get even more complicated. Indeed, we know from Benabou and Ok (2001) that if the income mobility process has a particular property (called *concavity in expectation*) that is often met in reality, then individuals who are currently poor may oppose redistribution because they (rationally) expect to move up in the income ladder in the future. Indeed, it is possible that this *prospect of upward mobility* (*POUM*) *hypothesis* may be strong enough to overturn the majority support of progressivity (even at the steady state of the underlying mobility process).

To wit, it seems at present that we do not have a general political theory of income taxation that provides theoretical support for the observed prevalence of progressive taxation schemes. Nevertheless, this is an area of active research, and it is not unreasonable to expect that lasting contributions on this theme will be made in the near future.

Optimal Tax Structures

Any discussion of progressive taxation would be incomplete without putting on record the large amount of work conducted towards the end of 20th century on the optimal design of income (and commodity) tax schemes. Indeed, after the seminal contribution of Mirrlees (1971), there has been an immense amount of work in this area, which has, however, slowed down significantly in more recent times. Roughly speaking, the canonical model of optimal income taxation works with a population with a given distribution of income earning ability and with a utility function having disposable income and effort as arguments, pre-tax income being a function of ability and effort. Individuals then act to maximize their individual utility subject to an income tax schedule which is to be determined to raise a given revenue while resulting in a maximum of utility for the population which is obtained as an aggregation of the individual utilities. Unfortunately, despite the promise of the early work on this topic – see, for instance, Sadka (1976), Stern (1976) and Seade (1977) – this model is found not to produce robust qualitative results. (See Stiglitz 1982, for a careful discussion of this issue.) One exception to this is the infamous *end-point theorem* that states that, when the distribution of skills has a known upper limit, the marginal tax rate should vanish at the level of income of the highest-income earners. (Informally speaking, the argument here is that there is no point to deterring the highest-income earner from earning the last dollar of her income, since if she does not earn it there will be no revenue from it.) This has the unsettling implication that an 'optimal' income tax is perforce non-progressive.

There is reason not to take this conclusion too seriously, however. First, simulations show that the end-point theorem is very local (cf. Tuomala 1990; Saez 2001). Second, the assumption that the constituent individuals are identical in all aspects but ability is a rather stringent requirement (which is key for the validity of the end-point theorem). Third, if there is uncertainty in the model that results in the expected income distribution having unbounded support, then the result fails. (See Haveman 1994, for a variety of other critical comments on optimal income taxation theory.) All in all, while it has duly brought the incentive problems into the forefront of public finance, and has provided a new means of evaluating the notion of equity–efficiency trade-off, it appears that the theory of optimal income (and consumption) taxation has not realized its promise in shedding light on the nature of progressive taxation, either normatively or positively.

Perhaps it is best to conclude our discussion, as Haveman (1994) does as well, with the words of Joseph Pechman, taken from his (posthumous) 1990 presidential address to the American Economic Association: 'Most people support tax progressivity on the ground that taxes should be levied in accordance with ability to pay, which is assumed to rise more than proportionately with income. Economists have . . . had trouble with the "ability to pay" concept . . . I believe that the person on the street is right and that we should continue to rely on the income tax to raise revenue in an equitable manner.'

See Also

- ▶ Optimal Taxation
- ▶ Tax Incidence

Bibliography

Atkinson, A. 1970. On the measurement of inequality. *Journal of Economic Theory* 2: 244–263.

Atkinson, A. 1994. The distribution of tax burden. In *Modern public finance*, ed. J. Quigley and E. Smolensky. Cambridge: Harvard University Press.

Benabou, R., and E.A. Ok. 2001. Social mobility and the demand for redistribution: The POUM hypothesis. *Quarterly Journal of Economics* 116: 447–487.

Berliant, M., and M. Gouveia. 1993. Equal sacrifice and incentive compatible income taxation. *Journal of Public Economics* 51: 219–240.

Besley, T., and S. Coate. 1997. An economic model of representative democracy. *Quarterly Journal of Economics* 112: 85–114.

Carbonell-Nicolau, O., and E. Klor. 2003. Representative democracy and marginal rate progressive income taxation. *Journal of Public Economics* 87: 2339–2366.

Carbonell-Nicolau, O., and E.A. Ok. 2007. Voting over income taxation. *Journal of Economic Theory* 134: 249–286.

Cukierman, A., and A. Meltzer. 1991. A political theory of progressive taxation. In *Political economy*, ed. A. Meltzer, A. Cukierman, and S. Richard. Oxford: Oxford University Press.

Eichhorn, W., H. Funke, and W. Richter. 1984. Tax progression and inequality of income distribution. *Journal of Mathematical Economics* 13: 127–131.

Fellman, J. 1976. The effects of transformations on Lorenz curves. *Econometrica* 44: 823–824.

Gouveia, M., and D. Oliver. 1996. Voting over flat taxes in an endowment economy. *Economics Letters* 50: 251–258.

Haveman, R. 1994. Optimal taxation and public policy. In *Modern public finance*, ed. J. Quigley and E. Smolensky. Cambridge: Harvard University Press.

Jakobsson, U. 1976. On the measurement of degree of progression. *Journal of Public Economics* 5: 161–168.
Kolm, S.-C. 1969. The optimal production of social justice. In *Public economics*, ed. J. Margolis and H. Guitton. London: Macmillan.
Kotlikoff, L., and L. Summers. 1989. Tax inecidence. In *Handbook of public economics*, ed. A. Auerbach and M. Feldstein, vol. 2. Amsterdam: North-Holland.
Lambert, P. 1999. Redistributional effects of progressive income taxes. In *Handbook of income inequality measurement*, ed. J. Silber. Boston: Kluwer.
Marhuenda, F., and I. Ortuño-Ortín. 1998. Income taxation, uncertainty and stability. *Journal of Public Economics* 67: 285–300.
Mirrlees, J. 1971. An exploration in the theory of optimum income taxation. *Review of Economic Studies* 38: 175–208.
Mitra, T., and E.A. Ok. 1996. Personal income taxation and the principle of equal sacrifice revisited. *International Economic Review* 37: 925–948.
Mitra, T., and E.A. Ok. 1997. On the equitability of progressive income taxation. *Journal of Economic Theory* 73: 316–334.
Musgrave, R., and P. Musgrave. 1989. *Public finance in theory and practice*. 5th ed. New York: McGraw Hill.
Ok, E.A. 1995. On the principle of equal sacrifice in income taxation. *Journal of Public Economics* 58: 453–468.
Osborne, M., and A. Slivinski. 1996. A model of political competition with citizen candidates. *Quarterly Journal of Economics* 111: 65–96.
Pechman, J. 1987. *Federal tax policy*. 5th ed. Washington, DC: Brookings Institution.
Pechman, J. 1990. The future of income tax. *American Economic Review* 80: 1–20.
Roberts, K. 1977. Voting over income tax schedules. *Journal of Public Economics* 8: 329–340.
Roemer, J. 1999. The democratic political economy of progressive income taxation. *Econometrica* 67: 1–19.
Romer, T. 1975. Individual welfare, majority voting, and the properties of a linear income tax. *Journal of Public Economics* 4: 163–185.
Rothschild, M., and J. Stiglitz. 1970. Increasing risk I: A definition. *Journal of Economic Theory* 2: 225–243.
Sadka, E. 1976. On income distribution, incentive effects and optimal income taxation. *Review of Economic Studies* 42: 261–268.
Saez, E. 2001. Using elasticities to derive optimal income tax rates. *Review of Economic Studies* 68: 205–229.
Seade, J. 1977. On the shape of optimal income tax schedules. *Journal of Public Economics* 7: 203–236.
Stern, N. 1976. On the specification of models of optimum income taxation. *Journal of Public Economics* 6: 123–162.
Stiglitz, J. 1982. Self-selection and Pareto efficient taxation. *Journal of Public Economics* 17: 213–240.
Thon, D. 1987. Redistributive properties of progressive taxation. *Mathematical Social Sciences* 14: 185–191.
Tuomala, M. 1990. *Optimal income tax and redistribution*. Oxford: Clarendon Press.
Vickrey, W. 1947. *Agenda for progressive taxation*. New York: Ronald Press.
Young, P. 1987. Progressive taxation and equal sacrifice principle. *Journal of Public Economics* 32: 203–214.
Young, P. 1990. Equal sacrifice and progressive taxation. *American Economic Review* 80: 253–266.

Project Evaluation

Partha Dasgupta

For a private commercial entrepreneur choice among alternative investment projects involves in principle a rather simple exercise. If he knows his own objectives, which would seem a reasonable assumption, all he has to do is to ascertain which projects satisfy his objectives best. For example, he may be interested solely in commercial profits; in which case commercial profitability – adjusted possibly for risk – is his criterion of choice.

The picture is a good deal more complex for a national planner. In choosing investment projects he has to ascertain which best satisfy the interests and objectives of the nation: his personal objectives are fairly unimportant. This is complex not merely because national objectives and interests are not easy to define, but also because the reading of these interests by different planners may well vary. If different planners pursue different national objectives the result may be unsatisfactory and conceivably disastrous.

There is another, equally important, reason why the problem is more complex for the national planner. To a private commercial entrepreneur market prices, when available, are hard bits of information. They enable him to estimate project profits. Not so for a national planner. To assess how an investment project affects national objectives the planner must have some understanding of how the economy will respond to the project; that is, what the project's impact is likely to be. For him market prices are often very soft data. He must have in addition some understanding of the technological and information constraints underlying the economy, in addition of course to the social and political constraints that impinge upon it.

The term 'project evaluation' refers to a branch of applied welfare economics concerned with just this class of issues: how should a national planner choose among alternative investment projects? It is part of a larger field of inquiry called) 'social cost-benefit analysis', which addresses the problem of evaluating all varieties of government activities, not just investment, with a view to providing a means of choosing among them. (See the pioneering work of Meade 1951, 1955; and also Arrow and Kurz 1970.) It should be emphasized though that the role of social cost-benefit analysis is not restricted to government use: the theory provides a conceptual framework with which any individual, as a citizen and social observer, can evaluate government action. Indeed, all concerned individuals perform social cost-benefit analysis continually.

Notional Prices

A particular approach to project evaluation consists of a systematic use of *notional prices* in valuing goods and services, estimating *notional profits* and then ranking investment projects in terms of such profits. These prices are notional because they are not necessarily equal to market prices. They are often called *accounting prices* (Tinbergen 1956; Little and Mirrlees 1969, 1974), and also *shadow prices* (Dasgupta et al. 1972). Mathematicians would call them *dual variables,* or *Lagrange multipliers.* They reflect the *social values* of goods and services measured in terms of national objectives. In a mixed economy the accounting price of a commodity or resource is the difference between its market price and the tax (or subsidy) that ought to be imposed on it. This last observation is useful to bear in mind because it reminds one that project evaluation cannot be thought about and conducted in a vacuum. Public investment choice must of necessity be integrated with the rest of government policy, in particular tax and trade policy (see Diamond and Mirrlees 1971, 1976; Mirrlees 1972; Dasgupta and Stiglitz 1972, 1974).

Selected History

The intellectual origins of the use of accounting prices lie in the brilliant work of Tjalling Koopmans on the possibilities of decentralizing the choice of optimal (or efficient) intertemporal production activities by the use of prices (see Koopmans 1951). In parallel was the establishment, due to Arrow (1951), of what is now called the Second Fundamental Theorem of Welfare Economics, which asserts that when preferences and production technologies are convex and closed, then under certain mild additional assumptions *any* Pareto-efficient preference allocation can be sustained in a decentralized environment if the government imposes appropriate lump-sump transfers among individuals and announces appropriate prices at which individuals and firms are instructed to trade. However, these contributions, and the immediate literature they inspired, were highly abstract and an attempt to make the entire approach operational was expounded by Tinbergen (1956).

There was in fact a parallel development. In market economies a *raison d'être* of government expenditure has been the supply of public goods, that is, commodities whose consumption among persons is non-rivalrous. In a seminal essay Samuelson (1954) revived Erik Lindahl's theory of public expenditure and established rules for the social cost-benefit analysis of government investment in the production of public goods. The key problem that was identified here lies in devising methods for estimating the social benefits from the consumption of public goods, for Samuelson noted that market data provide very soft information about these. (The theory of public goods was subsequently integrated with the economics of externalities by Arrow (1971). The recent literature on incentive compatibility and in particular preference revelation was much motivated by this problem. See e.g. Groves and Ledyard 1977; Green and Laffont 1977.)

A pioneering applied work on social cost-benefit analysis was a water-resource project studied by Robert Dorfman and his co-authors (see Dorfman et al. 1962). This did much to generate interest in the field. However, quite understandably the focus of attention shifted soon after to public investment in developing countries – where governments were busily engaged in producing ‘private goods’ like steel, cement, fertilizer and so forth. There was thus a need to integrate project evaluation within national plans and to ensure that national plans are influenced by the availability of desirable investment projects. This in turn meant that there was a need for a comprehensive analysis of accounting prices of various categories of goods and services in ‘second best economies’; that is, economies in which the government is forced to act in the presence not only of technological and resource constraints but also social, political and information constraints. This was presented in two treatises, Little and Mirrlees (1969) and Dasgupta et al. (1972), both aimed at developing countries, but differing sharply as regards government objectives and the extent to which project evaluators were seen to take constraints into account (see Dasgupta 1972). The decade of the 1970s saw an explosion of publications on theoretical and applied project evaluation based on the methods advocated by these treatises (e.g. Little and Mirrlees 1974; Little and Scott 1976; Squire and Van der Tak 1975; Hansen 1978; Helmers 1979). More recently both the theory and application of methods for estimating accounting prices have been extended to such categories of goods as environmental health benefits (Maler and Wyzga 1976; Freeman 1979), exhaustible resources (Dasgupta and Heal 1979) and renewable natural resources (Dasgupta 1982; Ahmad et al. 1984). We now present a formal model to illustrate the approach.

A Formal Model

Assume that there are $n + 1$ commodities labelled i or j, with $i, j = 0, 1, \ldots, n$. (Time can be introduced into the analysis by distinguishing goods by the date of their appearance and uncertainty by the states of nature contingent upon which they appear. See Debreu 1959.) Of these $n + 1$ goods $m + 1$ are non-traded ($i, j = 0, 1, \ldots, m$), the rest traded ($i, j = m + 1, \ldots, n$). For simplicity of exposition we take it that the country in question is small, so that we may as well normalize and set the border prices of traded goods equal to unity. We aggregate the government sector and denote by $\mathbf{z} = (z_0, z_1, \ldots, z_i, \ldots, z_n)$, the government net output vector and it will be understood that z_i is negative if i is a net input. Production possibilities in the government sector are represented by the set

$$G(\mathbf{z}) = G(z_0, z_1, \ldots, z_i, \ldots z_n) \leq 0.$$

It is best to aim at a general formulation of the problem and not be overly specific about the nature of the remaining sectors or indeed the extent of government control. As we are discussing project evaluation here we will certainly assume that the national planner chooses $\mathbf{z}$. However, we assume that the government can influence private consumption, private production and imports and exports only *indirectly.* Furthermore, we allow for the possibility that the private production, foreign trade and consumption sectors are non-competitive.

Let c_i be the aggregate consumption of commodity i. We assume that it is responsive to public production in a known way. This way we write as $c_i = c_i(\mathbf{z})$. (If there is uncertainty in this relationship the model can be extended in the obvious manner.) Now let y_i denote the net output of i in the private production sector. We suppose that it responds to public production $\mathbf{z}$ in a known way, $y_i(\mathbf{z})$. Private production is constrained by the relation, $y_0 = F(y_1, \ldots, y_n)$, where it should be emphasized that F is not necessarily a technological transformation curve but is, if the private sector is non-competitive, an amalgamation of technological possibilities and the industrial structure.

We denote by x_i ($i = m + 1, \ldots, n$) the net import of tradable i, and we let $x_i = x_i(\mathbf{z})$. Let R be the exogenously given endowment of foreign exchange, say aid. Finally we let social objectives be represented by the social welfare function W, defined directly on the aggregate consumption vector. Thus, $W = W(c_0(\mathbf{z}), c_1(\mathbf{z}), \ldots, c_i(\mathbf{z}), \ldots, c_n(\mathbf{z}))$. (Distributional issues can easily be incorporated into this, but at the expense of additional notation.)

The problem before the national planner is to:

Choose $\mathbf{z}$ so as to maximize $W(\mathbf{c}(\mathbf{z}))$ subject to

$$G(\mathbf{z}) \leq 0 \tag{1}$$

$$c_0(\mathbf{z}) \le F[y_1(\mathbf{z}), \ldots, y_n(\mathbf{z})] + z_0 \tag{2}$$

$$c_i(\mathbf{z}) \le y_i(\mathbf{z}) + z_i, \quad i = 1, \ldots, m \tag{3}$$

$$c_i(\mathbf{z}) \le y_i(\mathbf{z}) + z_i + x_i(\mathbf{z}), \quad i = m+1, \ldots, n \tag{4}$$

$$\sum_{i=m+1}^{n} x_i(\mathbf{z}) \le R. \tag{5}$$

Let non-negative numbers μ, λ_i $(i = 0, \ldots, n)$ and ρ be the Lagrange multipliers associated with the constraints (1), (2), (3), (4) and (5), respectively. It follows from the theory of non-linear programming that if there exist multiplier values μ^*, λ_i^* $(i = 0, \ldots, n)$ and ρ* such that they, in conjunction with the vector of public production, $\mathbf{z}^*$, constitute a saddle-point of the Lagrangean of the maximation problem, then **z*** is an optimal production vector: that is, a solution to the planning problem. It is now an easy matter to confirm that $\mathbf{z}^*$ must then satisfy the first-order conditions,

$$\begin{aligned} &\sum_{j=0}^{n}\left(\partial W/\partial c_j - \lambda_j^*\right)\partial c_j/\partial z_i \\ &+\sum_{j=1}^{n}\left(\lambda_0^*\partial F/\partial y_j + \lambda_j^*\right)\partial y_j/\partial z_i \\ &+\lambda_i^* + \sum_{j=m+1}^{n}\left(\lambda_j^* - \rho^*\right) \\ &\times\partial x_j/\partial z_i = \mu^*\partial G/\partial z_i, \quad \text{for}\, i = 0, \ldots, n. \end{aligned} \tag{6}$$

Equation (6) is fundamental to the theory and practice of project evaluation.

Let $i = 0$ be the numeraire good and define $p_i = \partial G/\partial z_i/\partial G/\partial z_0$ where the right-hand side is evaluated at $\mathbf{z}^*$. It follows that if G is a convex technology **z*** is a profit-maximizing production vector at prices $(1, p_1, \ldots, p_i, \ldots, p_n)$. These are the accounting prices to be used for project evaluation.

Certain special cases may now be mentioned. If the government has complete control over production, consumption and trade and if it wields these controls optimally and if the private production sector is competitive the economy attains a full-optimum allocation. (F in this case represents the private sector production possibility frontier.) According to the Second Fundamental Theorem of Welfare Economics this allocation satisfies the conditions,

$$\begin{aligned} \partial W/\partial c_j &= \lambda_j^* \qquad \text{for}\, j = 0, \ldots, n \\ \lambda_j^* &= -\lambda_0^*\partial F/\partial y_j \quad \text{for}\, j = 1, \ldots, n \end{aligned} \tag{7}$$

and

$$\lambda_j^* = \rho^* \quad \text{for}\, j = m+1, \ldots, n$$

From (6) and (7) we conclude that

$$p_i = \lambda_i^*/\lambda_0^* \quad \text{for}\, i = 0, \ldots, n \tag{8}$$

Conditions (7) and (8) tell us that accounting prices equal market prices and in particular that the shadow price of a tradable equals its border price. This is an undistorted economy.

A second important special case was studied by Diamond and Mirrlees (1971). They analysed optimal public policy in a world in which the private production and consumption sectors are competitive and in which, other than the production constraint G, the sole constraint facing the government is its inability to impose lump-sum transfers. (The rationale behind this inability could be incomplete information about private consumption needs.) Diamond and Mirrlees showed that at this particular ‘second-best’.

$$\sum_{j=0}^{n}\left(\partial W/\partial c_j - \lambda_j^*\right)\partial c_j/\partial z_i = 0, \quad \text{for}\, i = 0, \ldots, n$$

$$\lambda_j^* = -\lambda_0^*\partial F/\partial y_j, \quad \text{for}\, j = 1, \ldots, n$$

and

$$\lambda_j^* = \rho^*, \quad \text{for}\, j = m+1, \ldots, n. \tag{9}$$

It follows that $p_i = \lambda_i^*/\lambda_0^* \quad \text{for}\, i = 0, \ldots, n.$

The first condition in (9) circumscribes the set of optimal commodity taxes. The remaining conditions assert that, as in the Second Fundamental

Theorem, the government should seek economy-wide production efficiency (see also Dasgupta and Stiglitz (1972) and Little and Mirrlees (1974)).

In work parallel to this, Little and Mirrlees (1969, 1974), advocated the use of border prices of tradable commodites in project evaluation. This advocacy has been much debated (see e.g. the *Bulletin of the Oxford University Institute of Economics and Statistics,* February, 1972). It is a less stringent requirement than economy-wide production efficiency and thus might be expected to hold under a wider set of circumstances than envisaged in the Diamond–Mirrlees essay. This was precisely the point explored in Dasgupta and Stiglitz (1974), who located a wider class of circumstances in which

$$\sum_{j=0}^{n}\left(\partial W/\partial c_j-\lambda_j^*\right)\partial c_j/\partial z_i$$
$$+\sum_{j=1}^{n}\left(\lambda_0^*\partial F/\partial y_j+\lambda_j^*\right)\partial y_j/\partial z_i=0,\text{for}\, i=0,\ \ldots,n$$

and

$$\lambda_j^*=\rho^*\quad\text{for}\, j=m+1,\ \ldots,n$$

so that $p_i = 1$ for $i = m + 1, \ldots, n$: the Little–Mirrlees border-price rule.

Perturbation

In the above formulation accounting prices are obtained from an optimization exercise. A more restricted view of government capability sees it choosing investment projects sequentially on the basis of whether or not they *increase* social welfare (see Dasgupta et al. 1972). Let $\mathbf{z}'$ be the) 'current' state of public production. As it is feasible it must satisfy constraints (1), (2), (3), (4) and (5). An investment project is seen as a net addition to $\mathbf{z}'$ a perturbation, which we denote by Δz. For this to be feasible it must be that $\mathbf{z}'$ + Δz also satisfies constraints (1), (2), (3), (4) and (5). The criterion for choice is whether or not W increases as a consequence of accepting Δz. This, less ambitious, approach to project evaluation, originating in Meade (1951, 1955), is gaining much attention today, as it explicitly allows for limits on government rationality and will (see Hammond 1980; Blitzer et al. 1981).

Time and Social Discount Rates

It is not possible to discuss project evaluation without a mention of social rates of discount, unquestionably the most controversial matter in this field of discourse. We alter our notation somewhat and assume that there are k goods and services ($i = 1,\ldots, k$) at each date. Now let ($\mathbf{z}_1,\ldots, \mathbf{z}_t,\ldots$) be an intertemporal public production vector, where $\mathbf{z}_t$ is the k-vector of net outputs at date t. Now let $\mathbf{p}_t = (p_{1t},\ldots, p_{kt})$ be the accounting price k-vector at t ($t = 1, 2, \ldots$). It is a *present-value price vector.* Without loss of generality let $i = 1$ be the numeraire good at each date. Then the social rate of discount for the interval (t, $t + 1$) is defined as $(p_{1t}/p_{1t+1}) - 1$ (see Koopmans 1957).

Far too much has been written about social discount rates. (A comprehensive collection of essays is Lind 1982.) It is not that social discount rates are not important as an ingredient in project evaluation. They are. But then so are other accounting prices. Social rates of discount are the ratios of certain dated accounting prices. In general they are sensitive to the choice of social objectives and the economy-wide constraints, as is any other accounting price.

Increasing Returns

Convexity of the function G is a limiting assumption. Indeed, a *raison d'être* of the government engaging in the production of 'private' commodities is increasing-returns-to-scale. But this violates convexity. Clearly then project evaluation cannot be based exclusively on accounting prices. Even the classical marginal-cost pricing rule assumes that the government will provide a lump-sum subsidy to the public enterprise in order to cover its losses, and this argument is undertaken usually in a partial-equilibrium context. In a general case the matter is a great deal more complicated and no sure-fire rule for project evaluation would appear to be available. But the presence of scale-economies does not imply that accounting prices must be abandoned *in toto*. The contributions of Weitzman (1970) and Portes (1971) suggest that one ought to rely on accounting prices in the) 'convex' production sectors and quantity regulations in the) 'nonconvex' sectors. There is a great deal to be done in this field and we would expect it to be the area at which attention will now be directed.

See Also

- ▶ Cost–Benefit Analysis
- ▶ Development Economics
- ▶ Investment Decision Criteria
- ▶ Planning
- ▶ Shadow Pricing
- ▶ Tradable and Non-tradable Commodities

Bibliography

Ahmad, Y.J., P. Dasgupta, and K.G. Maler (eds.). 1984. *Environmental decision making*, vol. II. London: Hodder & Stoughton.

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. Berkeley: University of California.

Arrow, K.J. 1971. Political and economic estimation of social effects of externalities. In *Frontiers of quantitative economics*, vol. 1, ed. M. Intriligator. Amsterdam: North-Holland Publishing Company.

Arrow, K.J., and M. Kurz. 1970. *Public investment, the rate of return and optimal fiscal policy.* Baltimore: Johns Hopkins Press.

Blitzer, C., P. Dasgupta, and J.E. Stiglitz. 1981. Project appraisal and foreign exchange constraints. *Economic Journal* 91: 58–74.

Dasgupta, P. 1972. A comparative analysis of the UNIDO *guidelines* and the OECD *manual. Bulletin of the Oxford University Institute of Economics and Statistics* 34(1): 33–57.

Dasgupta, P. 1982. *The control of resources*. Oxford: Basil Blackwell.

Dasgupta, P., and G.M. Heal. 1979. *Economic theory and exhaustible resources*. Cambridge: Cambridge University Press.

Dasgupta, P., and J.E. Stiglitz. 1972. On optimal taxation and public production. *Review of Economic Studies* 39(1): 87–103.

Dasgupta, P., and J.E. Stiglitz. 1974. Benefit-cost analysis and trade policies. *Journal of Political Economy* 82: 1–33.

Dasgupta, P., S. Marglin, and A. Sen. 1972. *Guidelines for project evaluation*. New York: United Nations.

Debreu, G. 1959. *Theory of value*. New York: Wiley.

Diamond, P.A., and J.A. Mirrlees. 1971. Optimal taxation and public production: Parts I and II. *American Economic Review* 61: 8–27, 261–278.

Diamond, P.A., and J.A. Mirrlees. 1976. Private constant returns and public shadow prices. *Review of Economic Studies* 43: 41–47.

Dorfman, R., et al. 1962. *Design of water resource systems*. Cambridge, MA: Harvard University Press.

Freeman, A.M. 1979. *The benefits of environmental improvement: Theory and practice*. Baltimore: Johns Hopkins Press.

Green, J., and J.-J. Laffont. 1977. Characterization of satisfactory mechanisms for the revelation of preferences for public goods. *Econometrica* 45: 427–438.

Groves, T., and J. Ledyard. 1977. Optimal allocation of public goods: A solution to the free-rider problem. *Econometrica* 45: 783–810.

Hammond, P.J. 1980. Cost-benefit analysis as a planning procedure. In *Contemporary economic analysis*, vol. 2, ed. D.A. Currie and W. Peters. London: Croom Helm.

Hansen, R. 1978. *A guide to the guidelines*. New York: United Nations.

Helmers, F.L.C.H. 1979. *Project planning and income distribution*. London: Martinus Nijhoff.

Koopmans, T.C. (ed.). 1951. *Activity analysis of production and allocation*, Cowles Foundation monograph, no. 13. New York: Wiley.

Koopmans, T.C. 1957. *Three essays on the state of economic science*. New York: McGraw-Hill.

Lind, R. (ed.). 1982. *Discounting for time and risk in energy policy.* Baltimore: Johns Hopkins Press.

Little, I.M.D., and J.A. Mirrlees. 1969. *Manual of industrial project analysis in developing countries*. Paris: OECD.

Little, I.M.D., and J.A. Mirrlees. 1974. *Project appraisal and planning for developing countries*. London: Heinemann.

Little, I.M.D., and M.F.G. Scott (eds.). 1976. *Using shadow prices*. London: Heinemann.

Mäler, K.-G., and R.E. Wyzga. 1976. *Economic measurement of environmental damage*. Paris: OECD.

Meade, J.E. 1951. *The theory of international economic policy. Vol. I: The balance of payments*. London: Oxford University Press.

Meade, J.E. 1955. *Trade and welfare*. Oxford: Oxford University Press.

Mirrlees, J.A. 1972. On producer taxation. *Review of Economic Studies* 39(1): 105–111.

Portes, R. 1971. Decentralized planning procedures in centrally planned economies. *American Economic Review, Papers and Proceedings* 61: 422–429.

Samuelson, P.A. 1954. The pure theory of public expenditure. *Review of Economics and Statistics* 36: 387–389.

Squire, L., and H. Van der Tak. 1975. *Economic analysis of projects*. Baltimore: Johns Hopkins Press.

Tinbergen, J. 1956. *Economic policy: Principles and design*. Amsterdam: North-Holland.

Weitzman, M. 1970. Optimal growth with scale economies in the creation of overhead capital. *Review of Economic Studies* 37: 555–570.

Propensity Score

Jinyong Hahn

Abstract

Propensity score is an object often discussed in evaluation studies. It is defined as the conditional probability of treatment given covariates. It has attracted attention for its potential to control for the bias in the presence of high dimensional covariates.

Keywords

Asymptotic theory; Conditional probability; Covariate matching; Dimensionality; Evaluation research; Finite sample theory in econometrics; Matching; Nonparametric models and methods; Propensity score

JEL Classifications

C14

Propensity score is an object often discussed in evaluation studies. It is defined as the conditional probability of treatment given covariates. It has attracted attention for its potential to control for the bias in the presence of high dimensional covariates.

Evaluation research typically begins by comparing the treated group with the control group. For example, estimates of the effect of training programmes on earnings compare the earnings of those who receive training with a candidate control sample of untrained people. Because typically trainees are not chosen randomly, a simple comparison of the two groups may not provide a very accurate picture of what would have happened to the trainees had they not been trained. Under some conditions, such problems can be avoided by comparing the treated and the control groups with identical covariate values.

For more formal discussion, denote the covariate vector for person i by X_i, treatment status by D_i such that $D_i = 1$ if the ith person is treated and $D_i = 0$ otherwise, and define the conditional probability of treatment, or propensity score, as $p(X_i) \equiv \Pr[D_i = 1|X_i]$. Let Y_{1i} denote the potential, or counter-factual, outcome if the ith person receives the treatment, and let Y_{oi} denote potential earnings if he or she does not receive the treatment. Note that Y_{1i} is observed only when $D_i = 1$. Likewise, Y_{oi} is observed only when $D_i = 0$. This implies that $Y_{1I} - Y_{0i}$ is not observed by the researcher, and therefore the average treatment effects, which are defined to be $\beta = E[Y_{1i} - Y_{0i}]$, cannot be estimated by the sample analog of $E[Y_{1i} - Y_{0i}]$. Because D_i is not usually assigned randomly, a simple comparison of the two groups, that is, the sample analog of $E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 0]$, does not provide a consistent estimate of β, either. On the other hand, if (Y_{0i}, Y_{1i}) is independent of D_i given X_i, that is,

$$Y_{0i}, Y_{1i} \perp\perp D_i | X_i, \tag{1}$$

then the sample analog of

$$E\{E[Y_{1i}|X_i, D_i = 1] - E[Y_{0i}|X_i, D_i = 0]\}$$

will provide a consistent estimator for β. In other words, β can be consistently estimated by ‘matching’ the treated and the control groups with identical covariate values.

A problem that often arises in studies of this type is the need to control for continuously distributed and/or high-dimensional covariates. In many evaluation studies, the sample sizes are small, there are many covariates, and some of the covariates are continuous. A number of variations on exact covariate-matching schemes have been developed to deal with such situations. These typically involve approximate matching, or nonparametric smoothing, of some sort.

An alternative strategy to control for covariates begins with Rosenbaum and Rubin’s (1983) observation that bias can be eliminated by controlling for a scalarvalued function of the covariates, namely, the propensity score. Rosenbaum and Rubin's propensity-score theorem states that, if Eq. (1) is true, then it must also be true that conditioning on $p(X_i)$ eliminates selection bias, that is,

$$Y_{0i}, Y_{1i} \perp\perp D_i | p(X_i). \tag{2}$$

This implies that the β can be consistently estimated by the sample analog of

$$E\{E[Y_{1i}|p(X_i), D_i = 1] - E[Y_{0i}|p(X_i), D_i = 0]\}.$$

It is easier to estimate $E\{E[Y_{1i}|p(X_i), D_i = 1]$ than $E[Y_{1i}|X_i, D_i = 1]$ because the former requires the nonparametric regression of $Y_{1\mathrm{i}}$ on a scalar object $p(X_i)$ whereas the latter requires the nonparametric regression on a multi-dimensional object X_i. (Such difficulty is often called the curse of dimensionality in the nonparametrics literature.) The value of propensity score matching is in the 'dimension reduction' generated by regions where $p(X_i)$ is constant while $E[Y_{1i}|X_i]$ or $E[Y_{0i}|X_i]$ are not constant. The value of the propensity score is not clear, though, when the applied researcher does not have any information about the treatment assignment. Without such information, the propensity score needs to be estimated, which requires a nonparametric regression of D_i on X_i. Because this estimation suffers from the curse of dimensionality, the propensity score theorem seems to have little practical value when the propensity score itself needs to be estimated nonparametrically. On the other hand, it can be quite useful when applied researchers may have more information or are willing to make stronger assumptions about treatment assignment than about the relationship between covariates and outcomes. A number of empirical examples using the propensity score suggest that this approach works reasonably well (see, for example, Rosenbaum and Rubin 1984, 1985; Dehejia and Wahba 1999; Imbens et al. 2001; Heckman et al. 1998).

This evidence of practical utility notwithstanding, from an asymptotic theory point of view propensity-score-based estimators present a puzzle. Hahn (1998) shows that the propensity score is ancillary for estimates of average treatment effects, in the sense that knowledge of the propensity score does not lower the semiparametric efficiency bound for this parameter. Moreover, covariate matching is asymptotically efficient, that is, it attains the semiparametric efficiency bound, while propensity score matching does not. These results based on conventional asymptotic arguments seem to offer no justification for anything other than full control for covariates in estimation of average treatment effects.

The propensity score may still be a useful device. First, the propensity score may enhance finite sample efficiency. Angrist and Hahn (2004) use a non-standard asymptotic argument to point out that the traditional first-order asymptotic theory misses some of the subtleties in finite sample property, and observe that an estimator based on propensity score matching may be superior to the one based on covariate matching. They note, though, that the finite sample efficiency gain becomes smaller as the sample size grows, which is in accordance with the prediction from the traditional asymptotic theory. Second, when the estimated propensity scores are used as a weight in a certain way but not as a basis of matching, the estimator of the average treatment effects based on estimated propensity score is as asymptotically efficient as the covariate matching (Hirano et al. 2003).

See Also

- ▶ Finite Sample Econometrics
- ▶ Matching Estimators

Bibliography

Angrist, J., and J. Hahn. 2004. When to control for covariates? Panel-asymptotic results for estimates of treatment effects. *The Review of Economics and Statistics* 86: 58–72.

Dehejia, R., and S. Wahba. 1999. Propensity score matching methods for nonexperimental causal studies. *Journal of the American Statistical Association* 94: 1053–1062.

Hahn, J. 1998. On the role of the propensity score in the efficient estimation of average treatment effects. *Econometrica* 66: 315–332.
Heckman, J., H. Ichimura, and P. Todd. 1998. Matching as an econometric evaluation estimator. *Review of Economic Studies* 65: 261–294.
Hirano, K., G. Imbens, and G. Ridder. 2003. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* 71: 1161–1189.
Imbens, G., D. Rubin, and B. Sacerdote. 2001. Estimating the effect of unearned income on labor supply, earnings, savings and consumption: Evidence from a sample of lottery players. *American Economic Review* 91: 778–794.
Rosenbaum, P., and D. Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70: 41–55.
Rosenbaum, P., and D. Rubin. 1984. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* 79: 516–524.
Rosenbaum, P., and D. Rubin. 1985. Constructing a control group using multi-variate matching methods that include the propensity score. *American Statistician* 39: 33–38.

Propensity to Consume

John Eatwell

Keynes defined the propensity to consume as a functional relationship between the level of income and expenditure on consumption, and argued that ‘the amount that the community spends on consumption obviously depends (i) partly on the amount of its income, (ii) partly on the other objective attendant circumstances, and (iii) partly on the subjective needs and the psychological propensities and habits of the individuals composing it and the principles on which the income is divided between them’ (Keynes, 1936, pp. 90–91).

The determination of consumption by income, and, consequently, the determination of saving as a function of income, is the key proposition underpinning Keynes’s principle of effective demand. By means of this proposition he could transform Kahn’s multiplier (Kahn, 1931) into a theory of income and employment, in which the equality between investment and saving is ensured by appropriate adjustment in the level of income. Moreover, the fact that savings are a function of income whilst investment is an ‘autonomous’ expenditure (liberated from dependence on real saving by the revolving fund of finance) established the Keynesian proposition that in the aggregate investment *determines* saving.

An important characteristic of Keynes’s analysis of consumption is the absence of any behavioural role for the price mechanism. Relative prices might be assumed to have some impact on the composition of consumption, but the overall scale of consumption is a function only of the level, and perhaps the distribution, of income.

Keynes’s approach to the determination of consumption therefore stands in sharp contrast to the neoclassical analysis of consumption in which the composition *and* the scale of consumption is a function of relative prices, because income (or, more accurately, wealth) is determined by relative prices – i.e. by the valuation of individual endowments. If, in the neoclassical model, consumption is to be a function of income, then there must be some inhibition to the operation of the price mechanism which prevents the establishment of full Walrasian equilibrium. It was in this manner that Clower (1965) established the distinction between ‘effective’ and) ‘nominal’ demands, where the latter are associated with the Walrasian equilibrium, and the former are determined by income in a ‘rationed’ equilibrium (Benassy, 1975; Malinvaud, 1977).

The limited scope of Keynes’s critique of the neoclassical theory of employment, and his incorporation within his own analysis of some aspects of that theory in the form of the marginal efficiency of capital schedule, weakened his positive contribution, and resulted, ultimately, in the portrayal of the principle of effective demand as but another example of short-run market failure (see, for example, the papers in Harcourt, 1977). Yet, if an effective critique of the orthodox theory is deployed, then the principle of effective demand stands as the theory of the determination of output (Garegnani, 1964–65). At the core of that principle is Keynes’s) ‘fundamental psychological law’ – the propensity to consume.

See Also

- ▶ Consumer Expenditure
- ▶ Consumption function
- ▶ Life cycle hypothesis

Bibliography

Benassy, J.P. 1975. Neo-Keynesian disequilibrium theory in a monetary economy. *Review of Economic Studies* 42: 502–523.

Clower, R. 1965. The Keynesian counter-revolution: A theoretical appraisal. In *The theory of interest rates*, ed. F. Hahn and F. Brechling. London: Macmillan.

Garegnani, P. 1964–5. Note su consumi, investmenti, e domanda effectiva. *Economia Internazionale.* Reprinted in translation in *Keynes's economics and the theory of value and distribution,* ed. J. Eatwell and M. Milgate, London: Duckworth, 1983.

Kahn, R.F. 1931. The relation of home investment to unemployment. *Economic Journal* 41(June): 173–198.

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

Malinvaud, E. 1977. *The theory of unemployment reconsidered*. Oxford: Blackwell.

Property

Alan Ryan

Property rights are as fundamental to economics as scarcity and rationality. Unless some human agency has the right to control the use of whatever resource is in question nobody can set prices, and there will be no incentive for anyone to calculate costs of production. In much of their work, economists can, and do, take it for granted that everything of value (both tangible goods and intangible objects such as skills) has an owner, and that the owner's powers of control will correspond to the motivational assumptions of orthodox economic theory. That the free market and 'the liberal conception of ownership' (Honoré 1961) imply one another is obvious enough, and rightly allows most economists to feel free to leave the nature of ownership to others, while they tackle the intricacies of market interactions. Although Mill and Marx accused their contemporaries of discussing economics as if all the world had the legal institutions of the North Atlantic seaboard, the accusation is not wholly just – and both of them willingly exempted Smith in any case. For many purposes, the economy an investigator is concerned with can be assumed to have the legal background of the countries of the North Atlantic littoral. Nor have economists been reluctant to broaden their interest in property. Speculation about the possibilities of socialism, the analysis of the economics of slavery, inquiries into the agriculture of developing countries and very much more besides have all provoked investigations into the effects of particular systems of property rights. Defences of the free market based on individual private property (Buchanan 1985), explorations of the outlook for workers' cooperatives (Vanek 1970), assessment of the efficiency of American slavery (Fogel and Engerman 1974) and the literature provoked by Coase's demonstration of the irrelevance to overall welfare of the distribution of property rights are only a fraction of what economists have done.

No definition of ownership is wholly satisfactory for all purposes; 'the right of property is the right of dealing with things in the most absolute fashion the law allows', declares the French civil code, and it is echoed in many other codes. That seizes on two crucial things. First, it is not an infringement of my *ownership* of this knife that I may not stick it in your chest. The law allows nobody to stick a knife in anybody's chest, but whatever anyone may lawfully do with any knife, I (and nobody else) may do with this one. Second, the owner must have *all* the rights anyone can have over the things in question. The suggestion sometimes encountered in textbooks that ownership can be reduced to a 'right to an income' is inadequate, because it mistakes one element in ownership for the whole. Ownership certainly grounds the right to an income, but many rights to an income are grounded on something other than ownership, and ownership embraces other rights than the right to an income. The *code civile* is right to emphasize this, but does not deal with our intuition that if the law circumscribed too closely what an owner might do with his 'property', we might hesitate to call it his property at all. If, for instance, nobody could leave land to their children, or sell a freehold in it, or raise a mortgage on it, we should be doubtful whether individuals could be said to) 'own' it at all. The crucial element in the *'ius utendi et abutendi'* is the ultimate power of disposal. It is, therefore, not only a question of having all the rights the law allows, but of the law conferring on some person or institution the right of disposal. The same observation casts doubt on theories of so-called 'new property rights'. As did theorists of the 'new class', writers who have claimed to identify 'new property rights' have observed, rightly, that in a socialist society where there is supposedly no private ownership of the means of production, individuals in favoured positions may exert the same power as did capitalist owners of businesses in the past or have the same security of occupation and income as did those owners; similarly, those protected by trade unions in mixed economies may have the same security as those who purchased military commissions or government offices in the eighteenth century. Where they have gone wrong is in thinking of these as property rights. For they have simply ignored the question of who has the right of disposal. They may well be right to think that these new powers are as important as ownership and even that they confer on people powers similar to those which ownership confers. They are wrong to think that they amount to ownership.

Until the eighteenth century, and perhaps later, property was a central concern of political theorists. Plato began a long tradition in demanding that the rulers of the ideal republic should have no property. They should possess in common the common property of the republic, to separate their private interests from the public interest. He was not concerned, as St Paul was, to condemn avarice and urge men to set their hearts on the goods of the next world; rather he was concerned to avoid class warfare and to secure uncorrupt leadership. The lower classes were welcome to engage in their usual occupations and hang on to whatever few possessions those occupations yielded. Aristotle began an equally long tradition by observing that common property would not be regarded as) 'belonging to us all' but as 'belonging to no one'. Unless owned as private property, land and other resources would be neglected. But although private property was essential if people were to live in moderate comfort, Aristotle did not approve of the market; he complained that profit making was a distraction from the proper use of goods – which existed to be consumed not traded – and that lending money at interest was doubly wicked, because it was setting barren metal to breed. Ownership, and especially land ownership, existed in order to give the better sort of people the leisure to cultivate their talents and govern wisely.

Neither the Greeks nor the Romans had much use for the conception of individual rights which provides the framework of modern discussions of property. None the less, it was the Roman Law conception of ownership that bred modern theories of natural right and of a natural right of property, just as it was Roman thinking about practical politics that bred a rival habit of thought. This is the tradition of 'statecraft' and is exemplified in the work of Machiavelli and Harrington, and to a lesser extent in Hume and Smith. A crucial question to be asked of any system of property rights is whether it favours political stability and political liberty. The question is one of political sociology – what kind of property encourages public spirit in the citizen, and what kind encourages 'corruption' in the ruling class? The exemplary figure of the Roman farmer who kept his weapons over his fireplace and would fight for republican institutions against enemies from without and would-be tyrants from within haunts this tradition. Adam Smith, who is rightly thought of as the apostle of the modern economy, was equally taken with the ancient conviction that military valour, public spirit and free institutions were inconsistent with a wholly commercial economy. Small, independent farmers were the source of republican virtue. Their independence was not the same thing as modern individualism; they must be independent of the wealthy but they would not think of their land as theirs so much as their family's. In Machiavelli, the argument is entirely nostalgic; by the time of Harrington, there is more understanding of the impossibility of simply recreating the Roman republic; Hume and Smith saw that the modern, fluid, commercial world in which money is the great solvent of other forms of property cannot be escaped and cannot be wholly regretted. It is then an open question what balance of social forces can preserve freedom. It goes without saying that this 'political' conception of liberty with its roots in stable, landed property is not congenial to critics of the welfare state and socialism, who identify freedom with what Smith termed 'the simple system of natural liberty' or freedom of contract.

The 'statecraft' tradition does not enquire into the origins of property rights, nor into the justice of present distributions of property, but only into their political results. The natural law tradition (and its successors) is concerned with justice, and with what Locke called the 'original' of property. By 'original' he did not mean its historical origins but its moral logic. The crucial questions which this tradition faces are not sociological but moral – what grounds a valid claim to ownership; is there a conflict between the goals of property as an institution and the distribution of property rights in practice? From Locke, through Hume, Rousseau, Kant, Hegel, Mill and their successors, a variety of answers was offered, some of which had a clear tendency to justify the status quo, some of which, as in Rousseau on the one hand and Mill on the other, led to its rejection. Natural rights theories like Locke's held that each individual had a right to appropriate and use the unowned bounty of nature; the exercise of his natural liberty was enough to give him the ownership of it. Did it follow that the propertyless labourer in contemporary society had been cheated? Locke thought not; so long as he can earn a livelihood by his labour, he can 'appropriate' what he needs. But it does follow that owners who fail to provide employment commit an injustice. Rousseau held the same view, but complained that in practice the owners reduced the propertyless to near slavery and that even where they did not they corrupted them in other ways.

Hume and Mill offered a utilitarian justification for property. Unless there are rules of 'meum et tuum', there will be no efficient employment of the world's resources; as to what rules should govern ownership, that is a matter of expediency. But where Hume thought that expediency favoured custom and prescription even at the price of considerable inefficiency, Mill argued for positive governmental pressure by way of the law on property to promote efficiency on the one hand and the creation of an economy of producer cooperatives on the other.

In a very different idiom, Kant and Hegel also explained property as the expression of human freedom. Human beings, who alone possessed free-will, conferred value on the merely material objects they took into ownership. Without ownership, the world of mere material objects is inert, useless and of no value. But if property in some form or other is essential, the particular form is a matter for different governments to decide for themselves. Kant and Hegel were fierce enemies of the feudal hangovers which disfigured the German states of their day. Neither advocated complete laissez-faire, but since property expresses human sovereignty over nature, it must be open to any individual to acquire property by his own work. This rather romantic justification of property rights was turned on its head by Marx, when he declared that the irrationality of capitalism and its evident moral failings showed that so long as there was property at all, things would be sovereign over men and men would continue to suffer alienation. On the whole this argument has appealed to Marxist philosophers rather than Marxist economists; the economist feels he can analyse the consequences of different systems of public ownership, but has little to say about what a world without the very concept of property would be like.

In the mixed economies of the West, some kind of utilitarian justification of property is the 'common sense' of politics, even if most writers now acknowledge that the defence of private property needs to take into account questions of justice as well as questions of overall efficiency. John Rawls's insistence on appraising economic institutions from the standpoint of the representative least-favoured person (Rawls 1972) has captured the imagination of writers of a broadly, but uneasily utilitarian persuasion. The least abashed intellectual heirs of eighteenth and nineteenth-century utilitarianism are the defenders of the so-called ‘economic theory of property rights’. In this account the property rights characteristic of developed capitalist economies came into existence by an evolutionary process which allowed production to proceed in ever more efficient ways. The capitalist firm, say, exists because a system of property rights developed which allowed entrepreneurs to act swiftly and decisively. One implication is that a government which forced some other ownership pattern on the economy would find that evolutionary pressures would gradually reintroduce de facto capitalism and that only political repression could preserve socialism. The value of property rights lies in the pattern of resource management (in the widest sense) that they promote; what Marx condemned as) ‘bourgeois’ forms of ownership create the most efficient management. Critics complain that this suffers from the same defects as other doctrines of ‘the survival of the fittest’ – it takes its standard of fitness from the behaviour of the institutions it explains. But this is certainly one place where the discussion of property rights most vividly engages the concerns of lawyers, economists and philosophers alike.

See Also

- ▶ Entitlements

Bibliography

Alchian, A., and H. Demsetz. 1972. Production, information costs and economic organization. *American Economic Review* 62(5): 777–795.

Buchanan, A. 1985. *Ethics, efficiency and the market*. Totowa/Oxford: Rowman & Allanheld/Clarendon Press.

Buchanan, J.M. 1986. *Liberty, market and state*. Brighton: Wheatsheaf Books.

Coase, R. 1960. The problem of social cost. *Journal of Law and Economics* 3: 1–44.

Demsetz, H. 1967. Towards a theory of property rights. *American Economic Review,* Papers and Proceedings, 57: 347–359.

Fogel, R.W., and S.L. Engerman. 1974. *Time on the cross*. New York: Little, Brown.

Honoré, A.M. 1961. Ownership. In *Oxford essays in jurisprudence*, ed. A.G. Guest. London: Oxford University Press.

Rawls, J. 1972. *A theory of justice*. Oxford: Clarendon.

Vanek, J. 1970. *The general theory of labour-managed market economies*. Ithaca/London: Cornell University Press.

Property Law, Economics and

Dean Lueck

Abstract

This entry shows how the economics of property rights can be used to understand fundamental features of property law and related extra-legal institutions. It examines both the rationale for legal doctrine and the effects of legal doctrine regarding the exercise, enforcement, and transfer of rights. It also examines various property rights regimes including open access, private ownership, common property, and state property. Property law is understood as a system of societal rules designed to create incentives for people to maintain and invest in assets, which in turn leads to specialization and trade.

Keywords

Akerlof, G.: and land markets; Asymmetric information; And adverse selection; And servitudes; Average product rule; Coase Theorem; Common interest communities; Common property; Compensation; And investment decisions; Contracting: costs; Epstein, R.: and alienability of rights; Externalities; Common law responses; Neighbourhood; First possession; Land use: servitudes; Liability rules: vs. property rules; Negligence: vs strict liability; Open access property; Paradox of compensation; Posner, R.: and eminent domain; Private property; Property law: externalities; Property rights; Regimes; Origins; Alieneability; Property rules: vs liability rules; Public trust; Rule of capture; Servitudes; And asymmetric

P

information; Social norms: and property rights; State property; Strict liability: vs negligence; Takings: and US Constitution; Regulatory; Transaction costs; And Coase Theorem; Trespass: vs nuisance; Zoning

JEL Classifications

D23; D62; K11; K23

Property law is the body of court enforced rules that governs the establishment, use and transfer of rights to land and those assets attached to it such as air, minerals, water, and wildlife. In economic terms, property rights are defined as the (expected) ability of an economic agent to freely use an asset (Allen 1999; Barzel 1997; Lueck and Miceli 2007; Shavell 2004) and represent a social institution that creates incentives to use, to maintain, and to invest in assets. Property rights may or may not be enforced by courts; and because the actions of courts are costly legal rights are but a subset of economic property rights. In addition to law and regulations, property rights may be enforced by custom and norms (see, for example, Ellickson 1991) and by markets through repeated transactions.

Property Rights, Transaction Costs, and the Coase Theorem

Consider Coase's (1960) famous example of the rancher and farmer. The rancher's cattle stray onto the farmer's land causing crop damage. The rancher's profit, $\pi(h)$ and the amount of crop damage $d(h)$ are functions of the rancher's herd size h, so the first-best optimal herd size, h^* maximizes $\pi(h) - d(h)$ and h^* solves $\pi'(h) = d'(h)$. This is also the choice made by a single farmer-rancher, Coase's 'sole owner' case. If the rancher initially has the economic (and legal) right to impose crop damage without penalty, he would choose the herd size to maximize $\pi(h)$, adding cattle until $\pi'(h) = 0$, which implies $h^r > h^*$. The farmer would be willing to pay up to $d'(h)$, his marginal damage, for each steer that the farmer removes from the herd in order to avoid crop damage, while the rancher would accept any amount greater than his marginal profit, $\pi'(h)$.

If transaction costs are zero, the parties will instantly contract to reduce the herd to the efficient size. The farmer will purchase the rights to the straying cattle, and if the farmer had the initial rights the situation would be reversed: either way the outcome is first-best. This is the Coase Theorem: *When transaction costs are zero the allocation of resources will be efficient regardless of the initial assignment of property rights.* But transaction costs are not zero and thus property rights are not perfectly defined (Allen 1999; Barzel 1997; Lueck and Miceli 2007) so property law becomes important in defining rights and determining the allocation of assets. Indeed, Coase's (1960) discussion of nuisance law suggests an economic logic to the law in its assignment of property rights among various parties to these disputes.

Property Rights: Taxonomy and Models

Property law recognizes several fundamental property rights regimes: private property, open access, common property, and state property (Lueck and Miceli 2007). Property law also recognizes mixed regimes. Consider a fixed asset (such as a plot of land) used with a variable input (x) to produce a market output ($Y = f(x)$). If the input price is w, then the first-best use ($x^*(w)$) must maximize $R = f(x) - wx$ and satisfy $f'(x) = w$. The first-best value of the land is $V^* = \int_0^\infty R^*(x^*, t)e^{-rt}dt$, where r is the discount rate.

If there is 'open access' for n individuals, then output is $Y = f\left(\sum_{i=1}^n x_i\right)$ where x_i is the effort of the i^{th} individual, $f'(\cdot) > 0$ and $f'(\cdot) < 0$, and the opportunity cost of effort is w_i. Each person can only capture (and thus own) the output in proportion to his share of effort, so each solves:

$$\max_{x_i} R_i = f^i(x_i) - w_i x_i \ \text{ subject to } f^i = \left[x_i \Big/ \sum_{i=1}^n x_i\right] f\left(\sum_{i=1}^n x_i\right) \qquad (1)$$

On the assumption that users are homogeneous ($w_i = w_j.$ for all $i \neq j$), the Nash open access equilibrium is $x = x^{oa}(n,\ w_1,\ \ldots,\ w_n)$, which satisfies,

$$\begin{aligned}&(n-1/n)\left(f\left(\sum_{i=1}^{n} x_i\right)\Big/\sum_{i=1}^{n} x_i\right)\\&\quad-(1/n)f'\left(\sum_{i=1}^{n} x_i\right)\\&\quad=w_i,\quad i=1,\ldots n.\end{aligned}\tag{2}$$

In the limiting case as n → ∞, (2) becomes $f\left(\sum_{i=1}^{n} x_i\right)/\sum_{i=1}^{n} x_i = w$ which is the famous 'average product rule' (Gordon 1954; Cheung 1970; Brooks et al. 1999). The limiting case implies that rents are completely dissipated, or $\sum_{i=1}^{n} R_i = \sum_{i=1}^{n}\left[f^i(x^{oa}) - wx^{oa}\right] = 0$ and the present value of the asset is also zero, $V^{oa} = \int_0^\infty R(x^{oa}, t)e^{-rt}dt = 0$. With heterogeneous costs, the infra-marginal users earn rents and have incentives to maintain open access regime (Libecap 1989).

With private property the owner chooses $x^* < x^{oa}$ and generates $V^* > V^{oa} = 0$. Private ownership also creates incentives for optimal asset maintenance and investment (Bohn and Deacon 2000). Let future output be $Y_{t+1} = f(x_t)$, where x_t is current investment, available at a market wage of w. and the interest rate is r. The first-best use of the input $\left(x_t^*\right)$ must maximize $R = f(x_t)/(1+r) - w_t x_t$ and satisfy $f'(x_t)/(1+r) = w_t$. If $\pi \in [0,1]$ is the probability of expropriation (because of imperfect rights) of the future output, then an owner will maximize $R = f\ (x_t)[(1\ -\ \pi)/(1+r)] = w_t x_t$. The solution $\left(x_t^\pi < x_t^*\right)$ satisfies $f'(x_t)[(1-\pi)/(1+r)] = w_t$ and implies less than first-best investment. Pure open access means that no investor could claim future output ($\pi = 1$), so $x_t^{oa} = 0$, and the rent from investment also equals zero. This lack of incentive to invest is essentially the problem of the 'anti-commons' described by Heller (1998) and formalized by Buchanan and Yoon (2000).

Common property is exclusive ownership by a group and may arise out of explicit private contracting (for example, unitized oil reservoirs) or out of custom (for example, common pastures); it may have legal (for example, riparian water rights) or regulatory (for example, hunting regulations) bases that have implicit contractual origins. Common property is well documented for natural resource stocks in less developed economies (Bailey 1992; Ostrom 1990). It is also seen in modern 'common interest communities' (such as condominiums, homeowner's associations) where residents use quasi-governments to maintain common areas (such as pools, open space) and provide local public goods (Dwyer and Menell 1998). Contracting to form common property creates a group that can realize economies of enforcing exclusive rights. Equal sharing is a typical internal allocation rule; it avoids costs of measuring and enforcing individual use but still leads to overuse compared with first-best. With equal sharing rules a homogeneous membership maximizes the present value of a common property resource (Lueck 1994, 1995).

Governments own vast amounts of land, buildings, and capital equipment. State property rights are governed by administrative agencies, and the range of property rights regimes incorporates aspects of the three major types: private property, common property, and open access. State property rights commonly – and often severely – limit the transferability of rights, perhaps to limit the moral hazard incentives of agency bureaucrats. The relevant law for state property has its origins in common law (for example, mining on federal land is a first-possession rule) but is primarily governed by statutes and regulations, all shaped by bureaucrats, interest groups and politicians.

Real property regimes tend to mix the four fundamental types: open access, private property, common property and state property (Barzel 1982, 1997; Eggertsson 1990; Ellickson 1993; Kaplow and Shavell 1996; Merrill and Smith 2000; Rose 1998; Stake 1999), implicitly recognizing that assets are a collection of valuable attributes. A rancher's land is not typically completely private: the streams running through the property may be open access for fishing or recreation; the grass may be a lease from a federal agency with mineral rights held by yet another private party. Similar scenarios are found in residential and commercial real estate, and Bailey (1992) found a mixture of ownership regimes among aboriginal peoples. Smith's (2000) study of the common field system of medieval Europe is a rare study of the underlying economic logic of a mixed property regime.

Origin of Property Rights

In law and custom, first possession is the dominant method of establishing rights, be it to the *flow* of output from a *stock* or to the stock itself (Lueck 1995). Let $R(x(t))$ be the flow of benefits from an asset, where $x(t)$ is a variable input supplied at time t, r is the interest rate, and $g < r$ is the rate at which $R(t)$ grows over time. The first-best, full-information outcome is

$$V^{FB} = \int_{t=0}^{\infty} R(x^*(t))e^{-(r-g)t}dt, \qquad (3)$$

where $x^*(t)$ is the optimal input level and $t^* = 0$ since production begins immediately.

Under first possession the asset's first claimant obtains exclusive rights to the temporal flow of rents, $\int_0^\infty R^*(t)dt$. Since establishing a bona fide claim will be costly and because $g < r$. property rights to the asset will emerge as the value of the asset increases (Demsetz 1967). Along these lines an entire literature has developed to explain the 'evolution of property rights' or, more generally, the determinants – both temporal and cross section – of property rights regimes (Lueck and Miceli 2007; Rose 1998). This literature, mostly empirical, notes that property rights regimes can move in both directions (to and away from private property), that property rights regimes can move among mixed regimes, and that political and other institutions also shape the choice of property regimes.

Returning to first possession, a single claimant will choose the claiming time to maximize

$$V^S = \int_{t}^{\infty} \left[R(x^*(t))e^{-(r-g)t}dt\right] - Ce^{-rt}, \qquad (4)$$

where C is the cost of enforcing the claim and t is the time at which ownership of the stock (and the temporal flow of output) is established. The optimal time to establish ownership is when the present value of the asset's flow equals the present value of the opportunity cost of establishing rights at t^S, or $R^*e^{-(r-g)t^S} = rCe^{-rt^S}$. The asset value falls short of first-best, or $V^S < V^{FB}$, because the costs of establishing ownership delay ownership and production to t^S from $t = 0$.

First possession can dissipate value when there is unconstrained competition among homogenous claimants (Barzel 1968; Mortensen 1982). A competitive rush to claim rights causes ownership to be established at exactly the time t^R when the present value of the rental flow at t^R equals the present value of the entire costs of establishing ownership at t^R, or when $R^*e^{-(r-g)t^R}/(r-g) = Ce^{-rt^R}$. In this 'race equilibrium' rights are established at t^R, where $t^R < t^S$ since $t^R = (ln\,(r-g) + lnC - lnR)/g$ and $t^S = (lnr + lnC - lnR)/g$, and the rental stream is fully dissipated; or

$$V^R = \int_{t^R}^{\infty} \left[R(x^*(t))e^{-(r-g)t^R}dt\right] - Ce^{-rt^R} = 0. \qquad (5)$$

Heterogeneity among claimants can reduce, or eliminate, dissipation (Barzel 1994; Lueck 1995). If there are two competitors (i and j) with possession costs $C_i < C_j$, and neither party knows the other's costs, then i gains ownership just before j makes a claim, at $t^i = t^R - \varepsilon$, and earns rent equal to the present discounted value of his cost advantage. The key implication is that, as the differential between the two lowest cost claimants $(C_j - C_i)$ increases, the level of dissipation will decrease. With complete information there is no dissipation because only the low-cost claimant has a positive expected payoff in a race (Fudenberg et al. 1983; Harris and Vickers 1985).

If the costs of enforcing a claim to the asset are prohibitive, ownership may be established only by capturing or 'reducing to possession' a flow from the asset. The legal term 'rule of capture' describes this derivative of the rule of first possession. Wildlife and crude oil are the classic examples: ownership is established only when a hunter bags a pheasant or when a barrel of oil is brought to the surface. The stock itself (that is, the pheasant population or oil reservoir) remains unowned. The new 'race' is to claim the present flow $R(t)$ and leads to open access dissipation (Epstein 1986; Lueck 1995) since no one owns the asset's entire stream of flows, $\int_0^\infty R(t)dt$. The formal analysis is static rather than intertemporal as in the asset claim race, and is identical to the open access model developed above in eq. (1).

Property law implicitly recognizes the two potential paths of dissipation – racing and overexploitation – and is structured to limit such dissipation (Dharmapala and Pitchford 2002; Lueck 1995, 1998). Where first possession rules establish ownership in a resource stock, first possession tends to be defined so that valid claims are made at low cost and before dissipating races begin, thus exploiting claimant heterogeneity. Also, the transfer of rights to the resource is allowed, routinely reflecting security of ownership in the corpus. Where the rule of capture emerges (for example, oil and wildlife) access to the resource tends to be limited through legal, contractual or regulatory methods. As well, the transfer of rights to capturable flows tends to be restricted in order to limit overuse of the asset itself.

Externalities and Property Law: Nuisance, Trespass and Zoning

Externalities arise because property rights to at least some of the attributes of an asset will be imperfect and thus generate problems of open access or moral hazard. Land externalities are ubiquitous because any parcel (except an island) will have neighbouring owners and because related resources (for example, air, noise, minerals, water) do not tend to coincide with the surface ownership boundaries. Property law addresses externalities through doctrines of trespass, nuisance, servitudes, and through regulatory zoning.

Consider, à la Coase (1960), a railroad whose trains emit sparks that occasionally set fire to adjacent farmland. The number of trains is n_T and the number of farms is n_F, resulting in crop damage of $n_T n_F D(x, y)$, where D is the damage (reduced crop value per acre) each train causes, x is the cost of precaution per train, and y is the cost of precaution by each farmer. Assume $D_x < 0$, $D_y < 0$ $D_{xx} > 0$, and $D_{yy} > 0$. The marginal benefits are $b_T(n_T)$ and $b_F(n_F)$, where $b_j' < 0.j = T, F$. The total value of the two activities is

$$W = \int_0^{n_T} b_T(n_T)du + \int_0^{n_F} b_F(n_F)dz - [n_T n_F D(x, y) + n_T x + n_F y] \qquad (6)$$

If the numbers of trains and farms are fixed, as in tort models (Shavell 1980) that hold 'activity levels' fixed, the optimal precaution choices (x^*, y^*) that maximize (6) are $n_F D_x(x, y) + 1 = 0$ and $n_T D_y(x, y) + 1 = 0$. If the number of trains and farms (n_T, n_F) is endogenous, the resulting first-order conditions are $b_T(n_T) - [n_F D(x, y) + x] = 0$ and $b_F(n_F) - [n_T D(x, y) + y] = 0$.

Remedies for externalities can be viewed as a choice between 'property rules' and 'liability rules' (Calabresi and Melamed 1972; Polinsky 1980). Under property rules, rights holders can refuse any unwanted infringements of their rights, enforceable by injunctions (or criminal sanctions in the case of theft). Property rules thus form the legal basis for voluntary (market) exchange of rights. With liability rules, however, owners can only seek monetary compensation in the form of damages. Liability rules thus form the basis for court-ordered or non-consensual transactions. The choice between the two rules turns on transaction costs, particularly the costs of contracting, the costs of court adjudication, and legal administration. When contracting costs are relatively low, property rules are preferred because they ensure that all transactions are mutually beneficial. When contracting costs are high (for example, in public nuisance cases), property rules may prevent otherwise efficient transactions from occurring. Liability rules have an advantage because courts can force an efficient transfer. This advantage of liability rules must be weighed against the possibility of court error in setting damages, and, because liability rules require courts to establish the initial terms of a transaction by setting damages, the administrative costs of using this rule will likely be higher than under a property rule (Kaplow and Shavell 1996).

In the railroad–farmer case, if liability is strict the railroad must pay full compensation regardless of its level of precaution. Strict liability induces efficient precaution by the railroad, but farmers are fully compensated and thus have no incentive for precaution. Negligence, which holds the railroad liable for damages only if it takes less than the efficient level of abatement, will induce both parties to take efficient care. Neither rule, however, will achieve first-best railroad and farm activity levels. In general, liability rules cannot create first-best incentives because of the constraint that what one party pays the other must receive. This is an example of the *paradox of compensation* which is also found in tort law and contract law remedies (Cooter 1985). It can be avoided by 'decoupling' liability and compensation, or by using a contract or compensation mechanism that defines and enforces the optimal choices for both parties.

Trespass (for example, squatting, boundary encroachment) and nuisance (for example, air, water, noise pollution) doctrines are the primary common law responses to externalities. The primary remedy under trespass is an injunction, a property rule. The remedy under nuisance law is more complicated. A landowner can obtain relief only if the invasion is substantial, and even then he may have to be satisfied with money damages (a liability rule). If a landowner wishes the harm to be enjoined, he must meet the further legal standard of showing that the harm outweighs the benefit of the nuisance-creating activity. The trespass–nuisance distinction can be understood as a property–liability rule choice (Merrill 1998). Trespass ordinarily involves a small number of parties where the intruder is easily identifiable, so contracting costs tend to be low and property rules are likely optimal. Nuisance often involves large numbers or sources of harm that are difficult to identify, so liability rules are likely optimal.

Zoning is a common legal response to urban land externalities. The economic rationale for zoning is that 'similar land uses have no (or only small) external effects on each other whereas dissimilar land uses may have large effects' (White 1975), creating what the common law calls a 'public nuisance'. Ellickson (1973) argues that zoning may have administrative and enforcement costs that often exceed the saved 'nuisance costs'. A private alternative to zoning is the use of land use servitudes (for example, covenants, easements) that impose limits on what landowners can do with their property. Such restrictions are frequently observed in condominiums, homeowner associations, and other 'common interest communities' (Dwyer and Menell 1998; Hansmann 1991). The economic function of these restrictions is twofold: to overcome free rider problems in the provision of certain jointly consumed amenities; and to internalize neighbourhood and rental externalities.

Public Trust, Public Property and Public Use

The ancient doctrine of 'public trust' grants ownership of navigable rivers, shorelines, and the open sea to the public. It is judicially created common property, or sometimes open access. In its traditional application the public trust asset was a public good. When an asset is a public good, unrestricted access will not cause dissipation from overuse of the resource, but it could lead to underinvestment. When the resource has private good characteristics, unrestricted access can trigger the rule of capture and creates a classic open access problem, possibly causing resource degradation through overuse. Some courts have recently extended the doctrine into environmental assets, such as beaches, lakes, stream access and wildlife.

Large-scale projects like dams, railroads and highways often involve the assembly of a large contiguous parcel of land from relatively small and separately owned parcels. Developers face a potential holdout problem because, once assembly becomes public information, parcel owners might hold out for prices in excess of their true valuations, endangering completion of an otherwise efficient project. One solution is to force sales by replacing property rule protection of each owner's land with liability rule protection. This is the economic justification for the eminent domain power of the state (Posner 2003), which has common law origins. The ‘takings’ clause of the Fifth Amendment of the US Constitution explicitly grants such eminent domain power for ‘public use’ but requires ‘just compensation’, which courts have interpreted to mean ‘fair market value’. Since subjective value is part of the opportunity cost of a taking, failure to compensate for it potentially results in excessive acquisition of land by the government, though one study (Munch 1979) found that high-valued properties were overcompensated, while owners of low-valued properties were undercompensated.

A large literature has studied the link between compensation and investment decisions of landowners (Blume et al. 1984; Fischel and Shapiro 1988). Suppose there are many parcels, each worth $V(x)$ if the landowner makes an irreversible investment x, where $V' > 0$ and $V'' < 0$. The land also yields a public benefit of $B(y)$, where y is the number of parcels taken and $B' > 0$, $B'' < 0$. Compensation of $C(x)$ will be paid for each parcel taken, where $C(x) \geq 0$, $C' \geq 0$, and total compensation is $yC(x)$. Landowners choose x given the anticipated behaviour of the government and the compensation rule; then the government chooses y and pays $C(x)$. The first-best choices (x^*, y^*) maximize $B(y) + (1 - y)V(x) - x$, the sum of private and public benefits, and must satisfy $(1 - y)\,V'(x) - 1 = 0$ and $\mathrm{B}'(\mathrm{y}) - \mathrm{V}(\mathrm{x}) = 0$. If the taking is exogenous, y is fixed and the landowner will maximize $(1 - y)\,V(x) + yC(x) - x$, which must satisfy $(1 - y)\,V'(x) + yC'(x) - 1 = 0$ and also gives x^l as the solution. This means that compensation must be lump sum ($\mathrm{C}' = 0$) to ensure that $x^l = x^*$; a positive relationship between x and compensation creates over-investment moral hazard (another example of the paradox of compensation). Thus no compensation ($C(x) \equiv 0$ for all x) is actually efficient, although any lump sum rule is consistent with efficiency. The efficiency of zero compensation, however, depends on assumptions about government behaviour.

Government regulations often restrict land uses without depriving the owner of title (for example, zoning laws, environmental regulations). Historically, courts have granted broad powers to enact such regulations but, when a regulation becomes especially burdensome, the affected landowner may claim that a ‘regulatory taking’ has occurred and seek compensation. As above, the trade-off for regulatory takings concerns the efficiency of the land use decision on the one hand and the regulatory decision on the other. Miceli and Segerson (1994, 1996) propose the following compensation rule, where y is a landowner’s lost value from the regulation:

$$C = \begin{cases} 0 & \mathit{if} \quad y \leq y^* \\ V(x), & \mathit{if} \quad y > y^*. \end{cases} \tag{7}$$

Like a negligence rule in tort law, this rule requires full compensation if the government over-regulates ($y > y^*$) but requires no compensation otherwise ($y \leq y^*$). It also establishes a standard that is economically equivalent to the common law definition of a nuisance (an activity that is efficiently prohibited), and hence is consistent with the threshold for compensation implied by the nuisance exception.

Inalienability of Property Rights

Posner (2003, p. 75) notes, ‘the law should, in principle, make property rights freely transferable in order to allow resources to move to their most highly valued uses and to foster the optimal configuration of assets.’ Yet there are many legal restrictions that limit the alienability of property: body parts, children, voting, military service, cultural artifacts, endangered animal species, the right to freedom (laws against slavery), certain natural resources and state property.

The dominant economic reason for restrictions on alienability is that externalities can arise from transfers (Barzel 1997; Epstein 1985; Rose-Ackerman 1985; Posner 2003) if the rights to the assets are not well-defined with respect to the stock (and its stream of flows over time). This generates a rationale for limiting, even prohibiting, certain transfers of the claimed flows in order to protect the asset and its value. For example, the widespread prohibition on trade in wild game is likely to be such a case (Lueck 1989, 1998), though even here limits on markets can potentially deter the formation of property rights. Restrictions on the sale of children may have a similar rationale: a market for children (or game) would lead to 'poaching' of kids (or animals) for which property rights enforcement is extremely costly.

Another reason for restricting transfers is asymmetric information, particularly that leading to adverse selection (Rose-Ackerman 1998). Adverse selection can potentially dry up markets where product quality cannot be observed prior to purchase. Similar restrictions on the types of property servitudes allowed (such as limits on 'negative and in gross' easements) might be explained by reference to asymmetric information (Dnes and Lueck 2006). Legal scholars have argued that limitations on servitudes prevent 'clogging title' (Gray and Gray 2000). Consider the market for land of two types: fee simple (that is, unencumbered) and land encumbered with a servitude. Assume that only the seller knows whether the land is encumbered. Buyers do not have this information but know only that one-half of the land is encumbered. The value of an unencumbered plot is V^f, while the value of the encumbered plots is $V^s < V^f$. Given the information asymmetry, buyers will pay only the expected value of a plot, $EV = (V + V^f)/2 < V^f$. Following Akerlof (1970) and related literature, this means there will be no market equilibrium for the unencumbered plots; that is, only 'low-quality' encumbered plots will be present in the market. Institutions that provide information (such as title recording and registration systems) could eliminate asymmetry and even alter the law of property by allowing an expanded set of servitudes.

Summary

Economic analysis reveals a fundamental logic to the main doctrines and features of property law (Lueck and Miceli 2007). The observed structure of property rights and property law can be best understood as a system designed to create incentives for people to maintain and invest in assets, which in turn leads to specialization and trade. Among the most important remaining issues for study is a systematic analysis of how the law addresses the use and transfer of complex assets.

See Also

- ▶ Akerlof, George Arthur (Born 1940)
- ▶ contract Theory
- ▶ Law, Economic Analysis of
- ▶ Tragedy of the Commons

Bibliography

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

Allen, D. 1999. Transaction costs. In *Encyclopedia of law and economics*, ed. B. Bouckaert and G. DeGeest, Vol. I. Cheltenham: Edward Elgar.

Bailey, M. 1992. Approximate optimality of aboriginal property rights. *Journal of Law and Economics* 35: 183–198.

Barzel, Y. 1968. The optimal timing of innovations. *Review of Economics and Statistics* 50: 348–355.

Barzel, Y. 1982. Measurement cost and the organization of markets. *Journal of Law and Economics* 25: 27–48.

Barzel, Y. 1994. The capture of wealth by monopolists and the protection of property rights. *International Review of Law and Economics* 14: 393–409.

Barzel, Y. 1997. *Economic analysis of property rights*. 2nd ed. Cambridge: Cambridge University Press.

Besley, T. 1998. Investment incentives and property rights. In *The new Palgrave dictionary of economics and the law*, Vol. 2, ed. P. Newman. New York: Stockton Press.

Blume, L., D. Rubinfeld, and P. Shapiro. 1984. The taking of land: When should compensation be paid? *Quarterly Journal of Economics* 99: 71–92.

Bohn, H., and R.T. Deacon. 2000. Ownership risk, investment, and the use of natural resources. *American Economic Review* 90: 526–549.

Brooks, R., M. Murray, S. Salant, and J. Weise. 1999. When is the standard analysis of common property extraction under free access correct? *Journal of Political Economy* 107: 843–858.

Buchanan, J., and Y.J. Yoon. 2000. Symmetric tragedies: Commons and anticommons. *Journal of Law and Economics* 43: 1–14.

Calabresi, G., and A. Melamed. 1972. Property rules, liability rules, and inalienability: One view of the cathedral. *Harvard Law Review* 85: 1089–1128.

Cheung, S. 1970. The structure of a contract and the theory of a nonexclusive resource. *Journal of Law and Economics* 13: 49–70.

Coase, R. 1960. The problem of social cost. *Journal of Law and Economics* 3: 1–44.

Cooter, R. 1985. Unity in tort, contract, and property: The model of precaution. *California Law Review* 73: 1–51.

Demsetz, H. 1967. Toward a theory of property rights. *American Economic Review* 57: 347–359.

Dharmapala, D., and R. Pitchford. 2002. An economic analysis of 'riding to hounds': Pierson v. Post revisited. *Journal of Law, Economics, and Organization* 18: 39–66.

Dnes, A., and D. Lueck. 2006. An economic analysis of servitudes on land. Unpublished manuscript.

Dwyer, J., and P. Menell. 1998. *Property law and policy: A comparative institutional perspective*. Westbury: The Foundation Press.

Eggertsson, T. 1990. *Economic behavior and institutions*. Cambridge: Cambridge University Press.

Ellickson, R. 1973. Alternatives to zoning: Covenants, nuisance rules, and fines as land use controls. *University of Chicago Law Review* 40: 681–782.

Ellickson, R. 1991. *Order without law: How neighbors settle disputes*. Cambridge, MA: Harvard University Press.

Ellickson, R. 1993. Property in land. *Yale Law Journal* 102: 1315–1400.

Epstein, R. 1979. Possession as the root of title. *Georgia Law Review* 85: 1221–1243.

Epstein, R. 1985. Why restrain alienation? *Columbia Law Review* 85: 970–990.

Epstein, R. 1986. Past and future: The temporal dimension in the law of property. *Washington University Law Quarterly* 64: 667–722.

Fischel, W., and P. Shapiro. 1988. Takings, insurance, and Michelman: Comments on economic interpretations of 'just compensation' law. *Journal of Legal Studies* 17: 269–293.

Fudenberg, D., R. Gilbert, J. Sitglitz, and J. Tirole. 1983. Preemption, leapfrogging, and competition in patent races. *European Economic Review* 22: 3–31.

Gordon, H. 1954. The economic theory of a common-property resource: The fishery. *Journal of Political Economy* 62: 24–42.

Gray, K., and S. Gray. 2000. *Elements of land law*. 3rd ed. London: Butterworths.

Hansmann, H. 1991. Condominium and cooperative housing: Transactional efficiency, tax subsidies, and tenure choice. *Journal of Legal Studies* 20: 25–71.

Harris, C., and J. Vickers. 1985. Perfect equilibrium in a model of a race. *Review of Economic Studies* 52: 193–209.

Heller, M. 1998. The tragedy of the anti-commons: Property in transition from Marx to markets. *Harvard Law Review* 111: 621–688.

Kaplow, L., and S. Shavell. 1996. Property rules versus liability rules. *Harvard Law Review* 109: 713–790.

Libecap, G. 1989. *Contracting for property rights*. Cambridge: Cambridge University Press.

Lueck, D. 1989. The economic nature of wildlife law. *Journal of Legal Studies* 18: 291–324.

Lueck, D. 1994. Common property as an egalitarian share contract. *Journal of Economic Behavior and Organization* 25: 93–108.

Lueck, D. 1995. The rule of first possession and the design of the law. *Journal of Law and Economics* 38: 393–436.

Lueck, D. 1998. Wildlife law. In *The new Palgrave dictionary of economics and the law*, ed. P. Newman, Vol. 3. New York: Stockton Press.

Lueck, D., and T. Miceli. 2007. Property rights and property law. In *Handbook of law and economics*, ed. A. Polinsky and S. Shavell. Amsterdam: North-Holland.

Merrill, T. 1998. Trespass and nuisance. In *The new Palgrave dictionary of economics and the law*, ed. P. Newman, Vol. 3. New York: Stockton Press.

Merrill, T., and H. Smith. 2000. Optimal standardization in the law of property: The *numerus clausus* principle. *Yale Law Journal* 110: 1–70.

Miceli, T. 1997. *Economics of the law: Torts, contracts, property, litigation*. New York: Oxford University Press.

Miceli, T., and K. Segerson. 1994. Regulatory takings: When should compensation be paid? *Journal of Legal Studies* 23: 749–776.

Miceli, T., and K. Segerson. 1996. *Compensation for regulatory takings: An economic analysis with applications*. Greenwich: JAI Press.

Mortensen, D. 1982. Property rights and efficiency in mating, racing, and related games. *American Economic Review* 72: 968–979.

Munch, P. 1976. An economic analysis of eminent domain. *Journal of Political Economy* 84: 473–497.

Ostrom, E. 1990. *Governing the commons: The evolution of institutions for collective action*. New York: Cambridge University Press.

Polinsky, A. 1980. On the choice between property rules and liability rules. *Economic Inquiry* 18: 233–246.

Posner, R. 2003. *Economic analysis of law*. 6th ed. New York: Aspen Law and Business.

Rose, C. 1985. Possession as the origin of property. *University of Chicago Law Review* 53: 73–88.

Rose, C. 1986. The comedy of the commons, custom, commerce, and inherently public property. *University of Chicago Law Review* 53: 711–781.

Rose, C. 1998. Evolution of property rights. In *The new Palgrave dictionary of economics and the law*, ed. P. Newman, Vol. 2. New York: Stockton Press.

Rose-Ackerman, S. 1985. Inalienability and the theory of property rights. *Columbia Law Review* 85: 931–969.

Rose-Ackerman, S. 1998. Inalienability. In *The new Palgrave dictionary of economics and the law*, ed. P. Newman, Vol. 2. New York: Stockton Press.

Shavell, S. 1980. Strict liability versus negligence. *Journal of Legal Studies* 9: 1–25.

Shavell, S. 2004. *Foundations of economic analysis of law*. Cambridge, MA: Harvard University Press.

Smith, H. 2000. Semicommon property rights and scattering in the open fields. *Journal of Legal Studies* 29: 131–169.

Stake, J. 1999. Decomposition of property rights. In *Encyclopedia of law and economics*, ed. B. Bouckaert and G. DeGeest, Vol. 2. Cheltenham: Edward Elgar.

White, M. 1975. Fiscal zoning in fragmented metropolitan areas. In *Fiscal zoning and land use controls*, ed. E. Mills and W. Oates. Lexington: Lexington Books.

Property Rights

Armen A. Alchian

Keywords

Corporations; Exchange value; Firm; Theory of; Firm-specific resources; Fisheries; Government property rights; Joint production; Limited liability; Moral hazard; Mutual property rights; Ownership and control; Private property rights; Property rights; Quasirent; Specialization; Torts; Value

JEL Classifications

D0

Private Property Rights

A property right is a socially enforced right to select uses of an economic good. A Private property right is one assigned to a specific person and is alienable in exchange for similar rights over other goods. Its strength is measured by its probability and costs of enforcement which depend on the government, informal social actions, and prevailing ethical and moral norms. In simpler terms, no one may legally use or affect the physical circumstances of goods to which you have Private property rights without your approval or compensation. Under hypothetically perfect Private property rights none of my actions with my resources may affect the physical attributes of any other person's private property. For example, your Private property rights to your computer restrict my and everyone else's permissible behaviour with respect to your computer, and my Private property rights restrict you and everyone else with respect to whatever I own. It is important to note that it is the physical use and condition of a good that are protected from the action of others, not its exchange value.

Private property rights are assignments of rights to choose among inescapably incompatible uses. They are not contrived or imposed restrictions on the feasible uses, but assignments of exclusive rights to choose among such uses. To restrict me from growing corn on my land would be an imposed, or contrived, restriction denying some rights without transferring them to others. To deny me the right to grow corn on my land would restrict my feasible uses without enlarging anyone else's feasible physical uses. Contrived or unnecessary restrictions are not the basis of Private property rights. Also, because those restrictions typically are imposed against only some people, those who are not so restrained obtain a 'legal monopoly' in the activity from which others are unnecessarily restricted.

Under Private property rights any mutually agreed contractual terms are permissible, though not all are necessarily supported by governmental enforcement. To the extent that some contractual agreements are prohibited, Private property rights are denied. For example, it may be considered illegal to agree to work for over ten hours a day, regardless of how high a salary may be offered. Or it may be illegal to sell at a price above some politically selected limit. These restrictions reduce the strength of private property, market exchange and contracts as means of coordinating production and consumption and resolving conflicts of interest.

Economic Theory and Private Property Rights

A successful analytic formulation of Private property rights has resulted in an explanation of the method of directing and coordinating uses of economic resources in a private property system (that is, a capitalistic or a 'free enterprise' system). That analysis relies on convex preferences and two constraints: a production possibility and a private property exchange constraint, expressible biblically as 'Thou Shalt Not Steal', or mathematically, as the conservation of the exchange values of one's good.

For the decentralized coordination of productive specialization to work well, according to the well known principles of comparative advantage, in a society with diffused knowledge, people must have secure, alienable Private property rights in productive resources and products tradable at mutually agreeable prices at low costs of negotiating reliable contractual transactions. That system's ability to coordinate diffused information results in increased availability of more highly valued goods as well as of those becoming less costly to produce. The amount of rights to goods one is willing to trade, and in which Private property rights are held, is the measure of value; and that is not equivalent to an equal quantity of goods not held as private property (for example, government property). It probably would not be disputed that stronger Private property rights are more valuable than weaker rights, that is, a seller of a good would insist on larger amounts of a good with weaker Private property rights than if Private property rights to the goods were stronger.

Firms, Firm-Specific Resources and the Structure of Property Rights

Though Private property rights are extremely important in enabling greater realization of the gains from specialization in production, the partitionability, separability and alienability of Private property rights enables the organization of cooperative joint productive activity in the modern corporate firm. This less formally recognized, but nevertheless important, process of cooperative production relies heavily on partitioning and specialization in the components of Private property rights. Yet this method is often misinterpreted as unduly restrictive and debilitating to the effectiveness and social acceptability of Private property rights. To see the error, an understanding of the nature of the firm is necessary, especially in its corporate form, which accounts for an enormous portion of economic production. The 'firm', usually treated as an output-generating 'black box', is a contractually related collection of resources of various cooperating owners. Its distinctive source of enhanced productivity is 'team' productivity, wherein the product is not a sum of separable outputs each of which is attributable to specific cooperating inputs, but instead is a non-decomposable, non-attributable value produced by the group. Thus, for something produced jointly by several separately owned resources, it is not possible to identify or define how much of the final output value each resource could be said to produce separately. Instead, a marginal product value for each input is definable and measurable.

Whereas specialized production under comparative advantage and trade is directed in a decentralized process by market price and spot exchanges, productivity in the team, called the firm, relies on long-term, constraining contracts among owners who have invested in resources specialized to the group of inputs in that firm. In particular, some of the inputs are specialized to the team in that once they enter the firm their alternative (salvage) values become much lower than in the firm. They are called 'firm-specific'. In the firm, firm-specific inputs tend to be owned in common or else contracts among separate owners of the various inter-specific resources restrict their future options to those beneficial to that group of owners as a whole rather than to any individual. These contractual restrictions are designed to restrain opportunism and 'moral hazard' by individual owners, each seeking a portion of each other's firm-specific, expropriable composite quasi-rent. Taking only extremes for expository brevity, the other 'general' resources would lose no value if shifted elsewhere. A firm, then, is a group of firm-specific and some general inputs bound by constraining contracts, producing a non-decomposable end-product value. As a result, the activities and operation of the team will be most intensively controlled and monitored by the firm-specific input owners, who gain or lose the most from the success or failure of the 'firm'. In fact, they are typically considered the 'owners' or 'employers' or 'bosses' of the firm, though in reality the firm is a cooperating collection of resources owned by different people.

Firm-specific resources can be non-human. Professional firms – in law, architecture, medicine – are comprised of teams of people who would be less valuable elsewhere in other groups. They hire non-human general capital, for complex example building and equipment. The contract, which defines 'hiring', depends on the specificity and generality, not on human or non-human attributes nor on who is richer. Incidentally, 'industrial democracy' arrangements are rare, because the owners of more general resources have less interest in the firm than those of specific resources.

The Corporation and Specialization in Private Property Rights

In a corporation the resources owned by the stockholders are those the values of which are specific to the firm. The complexities in specialization in exercise of the components of property rights and the associated contractual restraints have led some people to believe that the corporation tends to insulate (for example, 'separate') decisions of use from the bearing of the consequences (that is, control from ownership) and thereby has undermined the capacity of a private property system to allocate resources to higher market value uses. For example, it has been argued that diffused stock ownership has so separated management and control of resources from 'ownership' that managers are able to act without sufficient regard to market values and the interests of the diffused stockholders. Adam Smith was among the first to propound that belief. Whatever the empirical validity, the logical analysis underlying those charges rests on misperceptions of the structure of Private property rights in the corporation and the nature of the competitive markets for control and ownership which tend to restrain such managers. What individual managers seek, and what those who survive are able successfully to do in the presence of competition for control, are very different things.

An advantage of the corporation is its pooling of sufficient wealth in firm-specific resources for large-scale operations. Pooling is enabled if shares of ownership are alienable private property, thereby permitting individuals to eliminate dependence of their time path of consumption on the temporal pattern of return from firm-specific investments. Alienability is enabled if the shares have limited liability, which frees each stockholder from dependence on the amount of wealth of every other stockholder. The resultant ability to tolerate anonymity, that is, disinterest in exactly who are the other shareholders, enables better market alienability.

When voluntary separability of decision authority over firm-specific resources from their market value consequences is added to alienability, the ability to specialize in managerial decisions and talent (control) without also having to bear the risk of all the value consequences, enables achievement of beneficial specialization in production and coordination of cooperative productivity. Specialization is not necessarily something that is confined to the production of different end products; it applies equally to different productive inputs or talents. Voluntary partitionability and alienability of the component rights enable advantageous specialization (sometimes called 'separation') in (*a*) exercise of rights to make decisions about uses of resources and (*b*) bearing the consequent market or exchange values. The former is sometimes called 'control' and the latter, 'ownership'. Separability enables the achievement of the gains from specialization in selecting and monitoring uses, evaluating the results, and bearing the risk of consequent future usefulness and value. Because different uses have different prospective probability distributions of outcomes, and because outcomes are differentially sensitive to monitoring the prior decisions, separability and alienability of the component rights permit gains from specialization in holding and exercising the partitionable rights.

Thus, the modern corporation relies on limited liability to enhance alienability and on partitionability of components of Private property rights in order to achieve gains from large-scale specialization in directing productive team activity and talents. Rather than destroying or undermining the effectiveness of Private property rights, the alleged 'separation' enables effective, productive 'specialization' in exercising Private property rights as methods of control and coordination.

Government Property Rights

It might be presumed that Government property rights in a democracy are similar to corporate property with diffused stockholdings and should yield similar results. The analogy would be apt if each voting citizen had a share of votes equivalent to one's share of the wealth in the community, and if a person could shift wealth among governments, as one can among different corporations. If, for example, one could buy and sell land (as assets capturing essentially most of the value of whatever the government does in that particular state) in several different governments and could vote in each in proportion to the value of that 'land' then government property would be closer to private property in its effects. But it is difficult to take that possibility seriously. The nature of government, public or communal property rights surely depends on the kind of government. Because these are so vaguely and indefinitely defined, attempts to deduce formally the consequences of resource allocation and behaviour under each have been hampered.

Non-Existent Property Rights

Not all resources are satisfactorily controlled by Private property rights. Air, water, electromagnetic radiation, noises and views are some examples. Water under my land flows to yours. Sounds and light from my land impinge on yours. Other forms of control are then designed, for example, political or social group decisions and actions, though these other forms are sometimes employed for ideological or political purposes, even where Private property rights already exist.

If these other forms permit open, free entry with every user sharing equally and obtaining the average return, use will be excessive. Extra uses will be made with an increased realized total value that is less than the cost added, that is, the social product value is not maximized. This occurs because the marginal yield is less than the average to each user, to which each user responds. So, use occurs to the point where the average yield is brought down to marginal cost, with the consequence that the marginal yield is less than the marginal cost – often exampled as excessive congestion on a public road or public park, or overfishing of communal, free access fishing areas. The classic 'communal property' implication that apples on the public apple tree are never allowed to ripen is an extreme example of the proposition that property rights, other than private, reduce conformity of resource uses to market revealed values. Alternatively, if communal property rights mean that incumbent users can block more users, the resource will be underutilized as incumbents maximize their individual yield, which is the average, not the marginal. This results in fewer users. Though more users or uses would lower the average value to the incumbents and hence dissuade a higher rate of use, the addition to the total group value (of the extra use) exceeds the extra costs. Examples are public, low tuition colleges that restrict entry to maximize the 'quality' of those who are educated – that is, to maximize the average yield of those admitted. Some labour unions (that is, teamsters) are examples of similar situations.

A mistaken inference commonly suggested by the example of fishermen who overfish unowned lakes is that independent sellers with open access to customers will 'over-congest' in product variety and advertising to catch customers, with unheeded costs borne by other sellers. If, for example, Pall Mall cigarettes attract some customers from Camel, the loss to Camel is the reduced value of Camel-specific resources, not its lost sales revenue. General resources will be released from making Camels for use elsewhere with no social loss. But Camel-specific resources fall in value by the extent to which Pall Mall's product is better or cheaper. Camel's loss is more than offset by the sum of Pall Mall's increased net income plus the transfer gain to customers from lower prices or better quality. The loss to Camel is not from new entry itself, but from its incorrect forecasts of its earlier investment value. It is presumed here that mistaken forecasts should not be protected by prohibiting the unexpected future improvements. This differs from the over-fishing case in that consumers, in contrast to fish, have property rights in what they pay and what they buy. If every fish had a separate owner or owned itself, none would allow it to be caught unless paid enough, and over-fishing would not occur. One owner of all the fish is unnecessary; it suffices that each fish (or potential customer) be owned by someone who can refuse to buy. (Of course, unless the lake were owned, the lake surface might be over-congested with too many fishermen, each fishing to a lesser area, even if the fish were owned.)

Ownership of tradable rights by customers is the feature that is missing in the over-fishing, over-congestion case. Because rights to (or 'of') the fish or whales need not be bought, over-fishing does not imply over-customering where customers own rights to what the competing sellers are seeking. Otherwise, customers could be caught like fish, wherein sellers would be competing both to (1) establish property rights over the customers and to (2) possess those rights. Costly redundant competition for initial establishment of rights could be avoided simply by establishing customers' rights to themselves, as is in fact done. If the preceding seems fanciful, replace 'fish' with people and the lake surface with streets on which taxi-drivers cruise for customers. Excessive costs will be incurred in competition for use of unowned, valuable resources, in this case, the streets.

Mutual Property Rights

'Mutual' forms of organization are used apparently in order to sustain the maximum average per member, or to reserve for the incumbent members any greater group value from more members. Mutual private property, a form that has barely been analysed, does not permit anonymous alienability of interests in what are otherwise Private property rights. A 'mutual' member can transfer its interest to other people only upon permission of the other mutually owning members or their agents. Fraternal, social and country clubs are examples. These activities have not typically been viably organized and their services sold, as for example, in restaurants and health and exercise gymnasia. The intragroup-specific resources are themselves the members (erstwhile customers) who interact and create their social utility. More members affect each incumbent's realized utility in two ways: by social compatibility and by congestion. An outside, separate owner interested in the maximum value of the organization, but not the maximum average-per-member, could threaten to sell more memberships which, although enabling a larger total social value with more members, would reduce the average value to the existing members. This is an example of the earlier analysed difference between maximizing the average yield per input rather than the total yield by admitting more members, who while they would be made better off than if not admitted nevertheless reduce the average value to the incumbent members. In addition, the ability of newcomers to compensate incumbents for any loss in the individual (average) value to incumbent members is restrained if the membership fee were to go instead to an outside owner of the club. To the extent that a pecuniary compensation, via an initiation fee, were paid to an outside owner and exceeded the reduction in their average individual and total group utility, newcomers would be admitted, and the outside owner would gain, but incumbent members would lose their composite quasi-rent of their interpersonal sociability. (It is not yet well understood why, aside from tax reasons, the mutual form occurs in savings and loans and insurance firms.)

Torts, Conditional and Unassigned Property Rights

Private property rights may exist in principle, but, quite sensibly, not be blindly and uncompromisingly enforced against all possible 'usurpers'. For example, situations arise in which someone's presumed Private property rights do not exclude an 'invader's' use. Accidental or emergency use of some other person's private property without prior permission constitutes an example, sometimes called a 'tort'. Another possibility is that the property rights are so ill-defined that whether a right has been usurped or already belonged to the alleged 'usurper' is unclear. For example, my newly planted tree may block the view from your land. But did you have a right to look across my land? If the rights to views (or light rays) were clearly defined and assigned, we could negotiate a price for preserving the view or my putting up a tree, depending upon which was more valuable to the both of us and with payment going to whoever proved to have the rights. Or, while sailing on a lake, to escape a sudden storm and save my boat and life, I use your dock without your prior permission. Did I violate any of your rights, or did your rights not include the right to exclude users in my predicament? If such emergency action is deemed appropriate, then rights to use of the dock are not all yours, as you may have thought. Whereas in the tree and view case, where a prior negotiation might have avoided a 'tort' (except that initially we did not agree about who had what rights), in the emergency use of the dock, prior negotiation was unfeasible. If prior negotiation is uneconomic, rights to that emergency use 'should' and will exist if that use is the most valuable use of the resource under the postulated circumstances. And compensation may or may not be required to the erstwhile 'owner'. The principle underlying such a legal principle seems straightforward and consistent with principles of efficient economic behaviour. It suffices for present purposes merely to call attention to this aspect of economic efficiency underlying the law.

See Also

- ▶ Clubs
- ▶ Coase Theorem
- ▶ Common Property Resources
- ▶ Fisheries
- ▶ Law, Economic Analysis of
- ▶ Property Law, Economics and
- ▶ Taking (Eminent Domain)

Property Taxation

George R. Zodrow

Abstract

Property taxation of both residential and nonresidential land and structures is the most common form of wealth taxation worldwide, and is often the revenue instrument of choice for local governments. Despite widespread use of the property tax and a voluminous academic literature examining the tax, its incidence and economic effects are still contentious issues, with the debate centring around whether the capital portion of the tax should be viewed as distorting the allocation of capital or as an efficient benefit tax or user charge for local public services.

Keywords

Assessment process; Benefit tax; Capital tax; Capitalization; Efficient allocation; Equity; Excise tax; Fiscal differential; Fiscal equalization; George, H.; General equilibrium; Harberger, A.C.; Hamilton, B.W.; Head tax; Housing market; Interurisdictional competition; Land tax; Land use zoning; Local government; Local public finance; Mieszkowski, P.; Partial equilibrium; Profits tax; Progressive and regressive taxation; Property taxation; Property tax limitations; Returns to capital; Tax incidence; Tiebout, C.; Tiebout model; Underprovision of public services; User fees

JEL Classifications

H2; H3; H4; H7

Property taxation of both residential and nonresidential land and structures – or 'real property' – is the most common form of wealth taxation worldwide, and is often the revenue instrument of choice for local governments. For example, in the United States property taxes account for over 70 per cent of local own-source tax revenues. Property tax liability is calculated as the product of the statutory rate and the assessed tax base, subject to a variety of adjustments, such as partial exemptions for primary residences and 'circuit breakers' designed to reduce tax burdens for certain groups, especially relatively poor elderly homeowners. Although vagaries in the assessment process have long been a source of inequity in the administration of the tax, recent advances in computer-based assessment practices have mitigated this problem. More recently, rapid growth in residential home values and the concomitant increase in property tax burdens and the share of total local taxes paid by homeowners have led to increasing dissatisfaction with the tax in some regions, culminating in the passage of numerous property tax limitation measures. In addition, concerns about the equity implications of financing primary and secondary public education with the property tax when the tax base, including non-residential property, is unequally distributed across school jurisdictions have led to reduced reliance on the tax as well as equalization mechanisms that redistribute property tax revenues across school districts (Anderson 1994).

The Incidence Debate: The Three Views of the Property Tax

The academic literature on the property tax – as reviewed by Mieszkowski and Zodrow (1989), Ladd (1998), Ross and Yinger (2000) and Netzer (2001) – has focused on the incidence and effects on the allocation of resources of the residential property tax. There is general agreement that the land component of the property tax is capitalized into land values, is borne by landowners at the time of the imposition of the tax, and – since land is immobile – does not distort the allocation of resources. Indeed, the efficiency advantages of taxing land values, coupled with a belief that most increases in land values reflect the benefits of public services, have led some observers, most prominently Henry George (1879), to advocate replacing property taxes with taxes on land values.

By comparison, the incidence and economic effects of property taxation of the capital or structures component of real property are among the most contentious issues in state and local public finance. Three views dominate the debate. The 'traditional' view dates back to Simon (1943) and Netzer (1966), who focused on the partial equilibrium effects of increasing the tax in a local housing market. From this perspective, one can make the standard 'open economy' assumption that the national return to capital is fixed. This in turn implies that local capital bears none of the local property tax, as capital in the long run migrates from the jurisdiction until the local after-tax return to capital equals the national value. The burden of the tax is thus borne by local factors and/or consumers, with the traditional view holding that the entire burden is borne by local housing consumers. The traditional view thus implies that the property tax inefficiently reduces the size of the local housing stock and that its burden is borne in proportion to housing consumption – and is thus somewhat regressive with respect to annual income but, more importantly, roughly proportional with respect to lifetime income.

A second popular theory is the 'benefit tax' view, developed by Hamilton (1975, 1976); Fischel (2001a, b) provides a recent discussion. This view is an extension of the renowned Tiebout (1956) model, which argues that consumer mobility ('voting with the feet') and inter–jurisdictional competition in the provision of local public services can ensure efficiency of resource allocation in the local public sector. Although Tiebout assumed the existence of benefit/head taxes, Hamilton extended the analysis by deriving conditions under which the property tax can be converted into the head tax assumed by Tiebout.

Specifically, following Tiebout, Hamilton assumes that individuals sort into local jurisdictions according to their demands for local public services, and that there are enough local tax-expenditure packages to accommodate all tastes. In addition, Hamilton (1975) assumes that local jurisdictions are homogeneous with respect to house values, and that there are enough jurisdictions to accommodate all desired housing and government service/tax packages. Finally, Hamilton assumes the existence of binding zoning constraints that established a minimum house value for each community. Under these circumstances, individuals are precluded from purchasing homes with a value below the minimum, and would never elect to pay taxes in excess of benefits received by purchasing a home with a value greater than the minimum; all individuals in a given community thus pay exactly the same property tax, which becomes a benefit tax.

Hamilton (1976) extends this model to the more realistic case in which house values within a community are heterogeneous. In this case, Hamilton assumes all communities are fully developed, effectively precluding any tax-induced changes in the housing stock, and that alternative communities which are homogeneous with respect to both demands for public services and housing are available. Under these circumstances, Hamilton shows that 'perfect capitalization' converts the property tax into a benefit tax, as a relatively expensive home sells at a discount equal to its 'fiscal differential' or the present value of all future taxes in excess of benefits received, while a relatively inexpensive home sells at a premium reflecting its fiscal differential, the present value of all future benefits in excess of future taxes. The implications of the benefit tax view are striking, as it implies that the property tax is effectively a non-distortionary user charge that has no direct effects on income distribution.

Finally, the 'capital tax' view (or 'new' view) of the property tax, developed by Mieszkowski (1972), subsequently extended by Zodrow and Mieszkowski (1983, 1986b), and reviewed in Zodrow (2001a, b), argues that the property tax is a distortionary tax on the local use of capital that results in a misallocation of the national capital stock across local jurisdictions. Mieszkowski (1972) stresses that earlier partial equilibrium analyses ignored the fact that the property tax is used by virtually all local jurisdictions and applies to a large fraction of the capital stock, including most non-residential capital. Adapting the Harberger (1962) general equilibrium model of tax incidence, he models the economy as having a fixed national capital stock and two types of local jurisdictions, characterized by 'high' tax rates and 'low' tax rates. In this context, Mieszkowski shows that property tax rates that exceed the national average drive capital out of high-tax jurisdictions into low-tax jurisdictions, with opposing effects occurring in relatively low tax jurisdictions. Property tax differentials thus result in an inefficient allocation of capital across jurisdictions. In addition, concern about the extent to which use of the property tax may drive capital out of a jurisdiction creates a tendency for local governments to under-provide public services (Zodrow and Mieszkowski 1986b; Wilson 1986). In terms of incidence, the 'average burden' of all of the property taxes imposed across the nation – known as the 'profits tax' effect of the tax – is borne by capital owners generally, implying that the tax is relatively progressive (with respect to annual income). In addition, Mieszkowski stresses that property tax differentials about the national average result in 'excise tax effects' in the form of housing and commodity price increases and wage and land price declines in relatively high-tax jurisdictions, with offsetting effects in relatively low-tax jurisdictions.

Differentiating Among the Three Views

Much of the subsequent literature has focused on reconciling or differentiating among these three views, and the issue of which view most accurately describes the effects of the property tax is still contentious. Matters are simplified somewhat because the traditional view has been shown to be a special case of the capital tax view. Specifically, the traditional view can be interpreted as a partial equilibrium analysis that focuses exclusively on the excise tax effects of the capital tax view, while neglecting its general equilibrium profits tax effects. Moreover, the traditional view that these excise tax effects are fully reflected in higher housing prices is accurate only under special circumstances; more generally, excise tax effects are borne in some combination by housing consumers and the owners of labour and land in the taxing jurisdiction (Wildasin 1986). In addition, the profits tax effect still obtains when one takes a general equilibrium perspective of the use of the property tax by a single small jurisdiction facing a perfectly elastic supply of capital. Specifically, although the capital outflow caused by an increase in the property tax by a small local jurisdiction depresses the overall return to capital only very slightly, this reduction affects a large capital stock; under certain circumstances, the overall reduction in national capital income precisely equals the revenue raised by the taxing jurisdiction, yielding the profits tax result (Zodrow and Mieszkowski 1983; Brown 1924; Bradford 1978). At the same time, the burden of a property tax increase in a single jurisdiction is borne entirely by local residents as higher prices or lower factor returns (with offsetting effects in all other jurisdictions). A critical implication is that even under the capital tax view there is a close link between local public services and the burden of the property tax, as the cost of financing local expenditures largely falls on local factor owners and consumers; thus, this interpretation of the capital tax view clearly has a strong benefit view flavour as local residents 'pay for what they get' in public services.

Nevertheless, the debate between proponents of the benefit tax view and the capital tax view is still far from resolved. The original Mieszkowski (1972) derivation, based on a model of national tax incidence, has been criticized for ignoring many of the features of local government service provision stressed by Tiebout and Hamilton. However, Zodrow and Mieszkowski (1986a) present an expanded derivation of the capital tax view that includes most of these aspects, including interjurisdictional competition, varying tastes for local public services, individuals sorting into differing communities according to tastes for local public services, and a simple form of land use zoning. They conclude that these factors thus do not distinguish between the two views; instead, the key factor in determining the incidence of the property tax is whether housing consumption can vary in response to the imposition of the tax. Moreover, although zoning requirements are pervasive, take a wide variety of forms, and can have a significant impact on property values (Fischel 1992), these facts do not demonstrate that zoning ordinances are sufficiently binding on housing consumption choices to ensure the validity of the benefit view (Rubinfeld 1987; Ross and Yinger 2000). In addition, although empirical evidence suggests that intrajurisdictional and intrajurisdictional capitalization of differences in property taxes and local expenditures is widespread (Oates 1969; Yinger et al. 1988; and Fischel 2001a, b, who concludes that evidence of 'capitalization is everywhere'), capitalization is consistent with both the assumption of fixed housing stocks that underlies the benefit tax view and the tax-induced reallocations of capital that underlie the capital tax view (Zodrow 2006); moreover, some observers have argued that in the long run capitalization is inconsistent with the benefit view (Ross and Yinger 2000). Finally, although some recent empirical tests are consistent with the capital tax view (Carroll and Yinger 1994; Wassmer 1993), these results are quite tentative. The primary empirical issue remaining to be resolved is whether the zoning restrictions or other mechanisms stressed by proponents of the benefit tax view are sufficiently binding to preclude the long-run adjustments in housing capital predicted by the capital tax view. This question promises to be a fertile topic for future research, which hopefully will help clarify the answer to the long-standing and critical questions of the incidence and economic effects of the property tax.

See Also

- ▶ Tax Incidence
- ▶ Taxation of Wealth

Bibliography

Anderson, J.E., ed. 1994. *Fiscal equalization for state and local government finance.* Westport: Praeger.

Bradford, D.F. 1978. Factor prices may be constant but factor returns are not. *Economics Letters* 1: 199–203.

Brown, H.G. 1924. *The economics of taxation*. New York: Holt.

Carroll, R., and J. Yinger. 1994. Is the property tax a benefit tax? The case of rental housing. *National Tax Journal* 47: 295–316.

Fischel, W.A. 1992. Property taxation and the Tiebout model: Evidence for the benefit view from voting and zoning. *Journal of Economic Literature* 30: 171–177.

Fischel, W.A. 2001a. Municipal corporations, homeowners, and the benefit view of the property tax. In *Property taxation and local government finance*, ed. W.E. Oates. Cambridge, MA: Lincoln Institute of Land Policy.

Fischel, W.A. 2001b. Homevoters, municipal corporate governance, and the benefit view of the property tax. *National Tax Journal* 54: 157–173.

George, H. 1879. *Progress and poverty.* New York: Dutton.

Hamilton, B.W. 1975. Zoning and property taxation in a system of local governments. *Urban Studies* 12: 205–211.

Hamilton, B.W. 1976. Capitalization of intrajurisdictional differences in local tax prices. *American Economic Review* 66: 743–753.

Harberger, A.C. 1962. The incidence of the corporate income tax. *Journal of Political Economy* 70: 215–240.
Ladd, H.F. 1998. *Local government tax and land use policies in the United States*. Northampton: Edward Elgar Publishing.
Mieszkowski, P. 1972. The property tax: An excise tax or a profits tax? *Journal of Public Economics* 1: 73–96.
Mieszkowski, P., and G.R. Zodrow. 1989. Taxation and the Tiebout model: The differential effects of head taxes, taxes on land rents, and property taxes. *Journal of Economic Literature* 27: 1098–1146.
Netzer, D. 1966. *Economics of the property tax*. Washington, DC: Brookings Institution Press.
Netzer, D. 2001. Local property taxation in theory and practice: Some reflections. In *Property taxation and local government finance*, ed. W.E. Oates. Cambridge, MA: Lincoln Institute of Land Policy.
Oates, W.E. 1969. The effects of property taxes and local public spending on property values: An empirical study of tax capitalization and the Tiebout hypothesis. *Journal of Political Economy* 77: 957–961.
Ross, S., and J. Yinger. 2000. Sorting and voting: A review of the literature on urban public finance. In *Handbook of regional and urban economics*, ed. P. Cheshire and E.S. Mills, vol. 3. Amsterdam: North-Holland.
Rubinfeld, D.L. 1987. The economics of the local public sector. In *Handbook of public economics*, ed. A.J. Auerbach and M.S. Feldstein. Amsterdam: North-Holland.
Simon, H.A. 1943. The incidence of a tax on urban real property. *Quarterly Journal of Economics* 59: 398–420.
Tiebout, C.M. 1956. A pure theory of local expenditures. *Journal of Political Economy* 64: 416–424.
Wassmer, R.W. 1993. Property taxation, property base, and property value: An empirical test of the ‘new view’. *National Tax Journal* 46: 135–160.
Wildasin, D.E. 1986. *Urban public finance*. New York: Harwood.
Wilson, J.D. 1986. A theory of inter-regional tax competition. *Journal of Urban Economics* 19: 296–315.
Yinger, J., H.S. Bloom, A. Boersch-Supan, and H.F. Ladd. 1988. *Property taxes and house values: The theory and estimation of intrajurisdictional property tax capitalization*. San Diego: Academic Press.
Zodrow, G.R. 2001a. Reflections on the new view and the benefit view of the property tax. In *Property taxation and local government finance*, ed. W.E. Oates. Cambridge, MA: Lincoln Institute of Land Policy.
Zodrow, G.R. 2001b. The property tax as a capital tax: A room with three views. *National Tax Journal* 54: 139–156.
Zodrow, G.R. 2006. *The property tax incidence debate and the mix of state and local finance of local public expenditures*, Working paper. Rice University.
Zodrow, G.R., and P. Mieszkowski. 1983. The incidence of the property tax: The benefit view vs. the new view. In *Local provision of public services: The Tiebout model after twenty-five years*, ed. G.R. Zodrow. New York: Academic Press.
Zodrow, G.R., and P. Mieszkowski. 1986a. The new view of the property tax: A reformulation. *Regional Science and Urban Economics* 16: 309–327.
Zodrow, G.R., and P. Mieszkowski. 1986b. Pigou, Tiebout, property taxation and the under-provision of local public goods. *Journal of Urban Economics* 19: 356–370.

Proportional Hazard Model

Jerry A. Hausman and Tiemen M. Woutersen

Abstract

This article reviews proportional hazard models and how the thinking about identification and estimation of these models has evolved since the mid-1970s.

Keywords

Baseline hazard; Convergence; Duration dependence; Duration measurements; Duration models; Gamma distribution; Heterogeneity; Mixed proportional hazard (MPH) model; Partial likelihood estimators; Proportional hazard models; Semiparametric estimation; Time-Varying regressors; Weibull model

JEL Classifications

C41

The estimation of duration models has been the subject of significant research in econometrics since the late 1970s. Cox (1972) proposed the use of proportional hazard models in biostatistics and they were soon adopted for use in economics. Since Lancaster (1979), it has been recognized among economists that it is important to account for unobserved heterogeneity in models for duration data. Failure to account for unobserved heterogeneity causes the estimated hazard rate to decrease more with the duration than the hazard rate of a randomly selected member of the population. Moreover, the estimated proportional effect of explanatory variables on the population hazard rate is smaller in absolute value than that on the hazard rate of the average population member and decreases with the duration. To account for unobserved heterogeneity Lancaster proposed a parametric mixed proportional hazard (MPH) model, a partial generalization of Cox's proportional hazard model, that specifies the hazard rate as the product of a regression function that captures the effect of observed explanatory variables, a baseline hazard that captures variation in the hazard over the spell, and a random variable that accounts for the omitted heterogeneity. In particular, Lancaster (1979) introduced the mixed proportional hazard model in which the hazard is a function of a regressor X unobserved heterogeneity v, and a function of time $\lambda(t)$,

$$\theta(t|X, v) = ve^{X\beta_0}\,\lambda(t). \qquad (1)$$

The function $\lambda(t)$ is often referred to as the baseline hazard and $v|X$ has a gamma distribution. The popularity of the mixed proportional hazard model is partly due to the fact that it nests two alternative explanations for the hazard $\theta(t|X)$ to be decreasing with time. In particular, estimating the mixed proportional hazard model gives the relative importance of the heterogeneity, v, and genuine duration dependence, $\lambda(t)$ (see Lancaster 1990, and Van den Berg 2001, for overviews). Lancaster (1979) uses functional form assumptions on $\lambda(t)$, which were not required by the Cox model, and distributional assumptions on v to identify the model. Examples by Lancaster and Nickell (1980) and Heckman and Singer (1984), however, show the sensitivity to these functional form and distributional assumptions. Thus, Lancaster's MPH model is fully parametric and from the outset questions were raised about the role of functional form and parametric assumptions in the distinction between unobserved heterogeneity and duration dependence. (Heckman 1991, gives an overview of attempts to make this distinction in duration and dynamic panel data models.) This question was resolved by Elbers and Ridder (1982), who showed that the MPH model is semi-parametrically identified if there is minimal variation in the regression function. A single indicator variable in the regression function suffices to recover the regression function, the baseline hazard, and the distribution of the unobserved component, provided that this distribution does not depend on the explanatory variables. Semi-parametric identification means that semi-parametric estimation is feasible, and a number of semi-parametric estimators for the MPH model have been proposed that progressively relaxed the parametric restrictions.

Nielsen et al. (1992) showed that the partial likelihood estimator of Cox (1972) can be generalized to the MPH model with gamma-distributed unobserved heterogeneity. Their estimator is semi-parametric because it uses parametric specifications of the regression function and the distribution of the unobserved heterogeneity. The estimator requires numerical integration of the order of the sample size, as originally discussed by Han and Hausman (1990), which further limits its usefulness and makes it impractical for most situations in econometrics. Heckman and Singer (1984) considered the non-parametric maximum likelihood estimator of the MPH model with a parametric baseline hazard and regression function. Using results of Kiefer and Wolfowitz (1956), they approximate the unobserved heterogeneity with a discrete mixture. The rate of convergence and the asymptotic distribution of this estimator are not known. As a result, these estimators that use discrete mixture with an increasing number of support points cannot be used to test hypotheses. Another estimator that does not require the specification of the unobserved heterogeneity distribution was suggested by Honoré (1990). This estimator assumes a Weibull baseline hazard and uses only very short durations to estimate the Weibull parameter.

Han and Hausman (1990) and Meyer (1990) propose an estimator that assumes that the baseline hazard is piecewise-constant, to permit flexibility, and that the heterogeneity has a gamma distribution. Both papers find that the hazard rate, conditional on heterogeneity, is non-monotonic so that the Weibull model cannot hold. Hausman and Woutersen (2005) present simulations and a theoretical result that show that using a nonparametric estimator of the baseline hazard with gamma heterogeneity yields inconsistent estimates for all parameters and functions if the true mixing distribution is not a gamma, which limits the usefulness of the Han–Hausman–Meyer approach. Thus, Hausman and Woutersen (2005) find it important to specify a model that does not require a parametric specification of the unobserved heterogeneity.

Horowitz (1999) was the first to propose an estimator that estimates both the baseline hazard and the distribution of the unobserved heterogeneity nonparametrically. His estimator is an adaptation of the semi-parametric estimator for a transformation model that he introduced in Horowitz (1996). In particular, if the regressors are constant over the duration, then the MPH model has a transformation model representation with the logarithm of the integrated baseline hazard as the dependent variable and a random error that is equal to the logarithm of a log standard exponential minus the logarithm of a positive random variable. In the transformation model the regression coefficients are identified only up to scale. As shown by Ridder (1990), the scale parameter is identified in the MPH model if the unobserved heterogeneity has a finite mean. Horowitz (1999) suggests an estimator of the scale parameter that is similar to Honoré's (1990) estimator of the Weibull parameter and is consistent if the finite mean assumption holds so that his approach allows estimation of the regression coefficients (not just up to scale). However, the Horowitz approach permits estimation of the regression coefficients only at a slow rate of convergence and it is not $N^{-1/2}$ consistent, where N is the sample size. The reason for the slower than $N^{-1/2}$ convergence is that the information matrix of the MPH model is singular under Horowitz assumptions (see Hahn 1994; Ishwaran 1996a). In particular, Horowitz (1999) assumes that the first three moments of the heterogeneity distribution exist, and Ishwaran (1996b) shows that the fastest possible rate of convergence is $N^{-2/5}$ for that case and Horowitz's (1999) estimator converges arbitrarily close to that rate. In other words, the slow rate of convergence is implied by the assumptions and is not a peculiarity of the estimator.

Subsequent research has focused on strengthening the assumptions of the MPH model so that $N^{-1/2}$ convergence is possible. Ridder and Woutersen (2003) derive a $N^{-1/2}$ consistent estimator for the MPH model by assuming that the baseline hazard rate is constant over a small interval, $\lambda(t)= \lambda$ for $0 \leq t \leq \varepsilon$ for any $\varepsilon > 0$ while allowing for a nonparametric baseline hazard function for $t > \varepsilon$. For parametric baseline hazards, Ridder and Woutersen (2003) assume that $\lim_{t\downarrow 0} \lambda(t)= \lambda$ for $0 < \lambda < \infty$ and derive another $N^{-1/2}$ consistent estimator. Hausman and Woutersen (2005) derive an estimator for the mixed proportional hazard model (with heterogeneity) that allows for a nonparametric baseline hazard and uses time-varying regressors. No parametric specification of the heterogeneity distribution or nonparametric estimation of the heterogeneity distribution is necessary. Intuitively, Hausman and Woutersen (2005) condition out the heterogeneity distribution, which makes it unnecessary to estimate it. Thus, they eliminate the problems that arise with the Lancaster (1979) approach to MPH models. In this model the baseline hazard rate is nonparametric, and the estimator of the integrated baseline hazard rate converges at the regular rate, $N^{-1/2}$ where N is the sample size. This convergence rate is the same rate as for a duration model without heterogeneity. The regressor parameters also converge at the regular rate. A nice feature of the estimator is that it allows the durations to be measured on a finite set of points. Such discrete measurement of durations is important in economics; for example, unemployment is often measured in weeks. In the case of discrete duration measurements, the estimator of the integrated baseline hazard converges only at this set of points, as would be expected.

It may be argued that the bias in the estimates of the regression coefficients is small if the estimates of the MPH model indicate that there is no significant unobserved heterogeneity. The problem with this argument is that estimates of the heterogeneity distribution are usually not very accurate. Given the results in Horowitz (1999), this finding should not come as a surprise. The simulation results in Baker and Melino (2000) show that it is empirically difficult to find evidence of unobserved heterogeneity, in particular if one chooses a flexible parametric representation of the baseline hazard. However, Han and Hausman (1990) and applications of their approach have found significant heterogeneity using a flexible approach to the baseline hazard. Bijwaard and Ridder (2002) find that the bias in the regression parameters is largely independent of the specification of the baseline hazard. Hence, failure to find significant unobserved heterogeneity should not lead to the conclusion that the bias due to correlation of the regressors and the unobservables that affect the hazard is small.

Because it is empirically difficult to recover the distribution of the unobserved heterogeneity, estimators that rely on estimation of this distribution may be unreliable. Therefore, it may be advisable to avoid estimating the unobserved heterogeneity distribution and the remainder of the MPH model simultaneously. Nevertheless, after estimating the baseline hazard and regression function, one can usually identify the mixing distribution. In particular, Horowitz (1999) uses the following equation to estimate the mixing distribution,

$$\ln\{\Lambda(T)\} + X\beta - \ln(Z) = -\ln(v)$$

where $\Lambda(T)$ and β can be estimated and the unobserved Z has an exponential distribution with mean one. Thus, Horowitz (1999) solves a deconvolution problem and the speed of convergence depends on the assumptions on the distribution of v.

A hazard model is a natural framework for time-varying regressors if a flow or a transition probability depends on a regressor that changes with time since a hazard model avoids the curse of dimensionality that would arise from interacting the regressors at each point in time with one another. A non-constructive identification proof for the duration model with time-varying regressors can be produced using techniques similar to Honoré (1993b), and Honoré (1993a) gives such a proof. (A non-constructive identification proof is an identification proof that does not suggest an estimator.) In particular, Honoré (1993a) does not assume that the mean of the heterogeneity distribution is finite (nor does Honoré 1993a, assume a tail condition as in Heckman and Singer 1984). Ridder and Woutersen (2003) argue that it is precisely the finite mean assumption that makes the identification of Elbers and Ridder (1982) ‘weak’ in the sense that the model of Elbers and Ridder (1982) cannot be estimated at rate $N^{-1/2}$. As in Honoré (1993a), Hausman and Woutersen (2005) do not need the finite mean $N^{-1/2}$ assumption which gives an intuitive explanation of why Hausman and Woutersen (2005) can estimate the model at rate $N^{-1/2}$.

Bibliography

Baker, M., and A. Melino. 2000. Duration dependence and nonparametric heterogeneity: A Monte Carlo study. *Journal of Econometrics* 96: 357–393.

Bijwaard, G., and G. Ridder. 2002. Efficient estimation of the semi-parametric mixed proportional hazard model. Working paper.

Cox, D. 1972. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society B* 34: 187–220.

Elbers, C., and G. Ridder. 1982. True and spurious duration dependence: The identifiability of the proportional hazard model. *Review of Economic Studies* 49: 402–409.

Hahn, J. 1994. The efficiency bound of the mixed proportional hazard model. *Review of Economic Studies* 61: 607–629.

Han, A., and J. Hausman. 1990. Flexible parametric estimation of duration and competing risk models. *Journal of Applied Econometrics* 6: 1–28.

Hausman, J., and T. M. Woutersen. 2005. Estimating a semi-parametric duration model without specifying heterogeneity. Working paper, UCL: CeMMAP, Institute for Fiscal Studies.

Heckman, J. 1991. Identifying the hand of the past: Distinguishing state dependence from heterogeneity. *American Economic Review* 81: 75–79.

Heckman, J., and B. Singer. 1984. A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica* 52: 271–320.

Honoré, B. 1990. Simple estimation of a duration model with unobserved heterogeneity. *Econometrica* 58: 453–473.

Honoré, B. 1993a. Identification results for duration models with multiple spells or time-varying regressors. Northwestern working paper.

Honoré, B. 1993b. Identification results for duration models with multiple spells. *Review of Economic Studies* 60: 241–246.

Horowitz, J. 1996. Semiparametric estimation of a regression model with an unknown transformation of the dependent variable. *Econometrica* 64: 103–107.
Horowitz, J. 1999. Semiparametric estimation of a proportional hazard model with unobserved heterogeneity. *Econometrica* 67: 1001–1028.
Ishwaran, H. 1996a. Identifiability and rates of estimation for scale parameters in location mixture models. *Annals of Statistics* 24: 1560–1571.
Ishwaran, H. 1996b. Uniform rates of estimation in the semiparametric Weibull mixture model. *Annals of Statistics* 24: 1572–1585.
Kiefer, J., and J. Wolfowitz. 1956. Consistency of maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics* 27: 887–906.
Lancaster, T. 1979. Econometric methods for the duration of unemployment. *Econometrica* 47: 939–956.
Lancaster, T. 1990. *The econometric analysis of transition data*. Cambridge: Cambridge University Press.
Lancaster, T., and S. Nickell. 1980. The analysis of re-employment probabilities for the unemployed. *Journal of the Royal Statistical Society, A* 143: 141–165.
Meyer, B. 1990. Unemployment insurance and unemployment spells. *Econometrica* 58: 757–782.
Nielsen, G., R. Gill, P. Andersen, and T. Sørensen. 1992. A counting process approach to maximum likelihood estimation in frailty models. *Scandanavian Journal of Statistics* 19: 25–43.
Ridder, G. 1990. The non-parametric identification of generalized accelerated failure time models. *Review of Economic Studies* 57: 167–182.
Ridder, G., and T. Woutersen. 2003. The singularity of the information matrix of the mixed proportional hazard model. *Econometrica* 71: 1579–1589.
Van den Berg, G. 2001. Duration models: Specification, identification, and multiple duration. In *Handbook of econometrics*, ed. J. Heckman and E. Leamer, Vol. 5. Amsterdam: North-Holland.

Prospect Theory

Graham Loomes

Abstract

Prospect theory sought to provide a descriptive model of risky choice which could accommodate a number of seemingly systematic violations of conventional ‘expected utility’ analysis. Although there are phenomena which the model cannot explain (even in its later ‘cumulative’ form), it constitutes a landmark in the development of alternative theories which have modified standard theory and/or have tried to incorporate psychological factors into decision theories.

Keywords

Allais paradox; Behavioural economics; Choice; Common consequence effect; Common ratio effect; Cumulative prospect theory; Decision weights; Expected utility hypothesis; Isolation effect; Kahneman, D.; Probability distribution; Prospect theory; Rational behaviour; Reference point; Reflection effect; Stochastic dominance; Subjective probability; Transitivity; Tversky, A

JEL Classifications

C9

Prospect theory (PT) was developed by psychologists Daniel Kahneman and Amos Tversky to try to account for a number of patterns of response to risky choices which departed systematically from the conventional wisdom about rational decision making in the form of von Neumann and Morgenstern’s (1944) expected utility (EU) hypothesis.

Kahneman and Tversky’s (1979) paper ‘Prospect Theory: An Analysis of Decision Under Risk’ has proved to be enormously influential. According to Kim, Morse and Zingales (2006), it is the second most frequently cited paper published in economics journals since 1970, with more than 4,000 citations in the 25 years since its publication. It provided a major stimulus to the development of a number of other ‘non-expected-utility’ theories in the 1980s and 1990s – see Starmer (2000) for a survey and review. It has also inspired much work in behavioural economics and in economic and psychological experiments exploring individual decision making under risk and uncertainty.

The following subsections consider what PT set out to do and how it did it. There then follows a discussion of the importance of the theory as well as its possible limitations.

Background

In the 1950s and 1960s, evidence had begun to accumulate which suggested that EU failed as a general *descriptive* model of risky choice. Two of the most influential 'paradoxes' had been identified by Maurice Allais in the early 1950s (see Allais 1953). These were renamed by Kahneman and Tversky (henceforth K&T) and are now widely known as the 'common ratio effect' and the 'common consequence effect'. Briefly, they are as follows, starting with the common ratio effect.

Consider the choice between two 'prospects' A and B where A offers a sum of money x with probability p (and 0 with probability $1 - p$) while B offers a smaller sum y with a larger probability q (and 0 with probability $1 - q$). An extreme form of this might involve setting $q = 1$, so that B offers the certainty of y: in an example used by K&T, B offered the certainty of 3,000 Israeli pounds while A offered a 0.8 chance of 4,000 (and a 0.2 chance of 0). EU does not predict which of A and B an individual will choose – that depends on the individual's personal tastes concerning risk – but what the independence axiom of EU *does* entail is that, if p and q are scaled down by the same factor so that the ratio of 'winning' probabilities is maintained, the preference between the scaled-down prospects will be consistent with the preference between A and B.

So – to continue the example used by K&T – suppose that both p and q are scaled down to a quarter of their original values, generating prospects C and D, where C offers a 0.2 chance of 4,000 and a 0.8 chance of 0, while D offers a 0.25 chance of 3,000 and a 0.75 chance of 0. Then EU entails that anyone who prefers A over B should also prefer C over D, and vice versa. However, the common ratio effect form of the Allais paradox is manifested when a substantial proportion of those who choose the safer option B in the first case switch to the riskier prospect C in the second case, while the combination of choosing A in the first case and D in the second case is relatively rare.

In the above case, the scaling down operated on the *magnitudes* of the winning probabilities, while *maintaining the ratio* between them. Another way of manipulating the prospects could work in terms of replacing some probability of a particular sum common to both prospects by the same probability of a different sum. Consider another example used by K&T. This time, E offers 2,500 with probability 0.33, 2,400 with probability 0.66 and 0 with probability 0.01, while F offers 2,400 with certainty. Now, for both prospects, replace the 0.66 probability of 2,400 by a 0.66 probability of 0: this transforms E into a prospect G which offers a 0.33 chance of 2,500 and a 0.67 chance of 0, and transforms F into a prospect H which offers a 0.34 chance of 2,400 and a 0.66 chance of 0. Once again, EU entails that individuals should either choose E in the first case and G in the second, or else they should choose F and H. However, the common consequence effect form of Allais paradox involves many more individuals switching from safer to riskier (that is, choosing F and G) than switch from riskier to safer (that is, choose E and H).

In addition to the common ratio and common consequence effects, two other 'effects' were influential in the formulation of PT. One of these is the 'isolation effect'. Consider again the 'scaled-down' pair of prospects from the common ratio example. In the way they were presented there, C offered a 0.2 chance of 4,000 together with a 0.8 chance of 0, while D offered a 0.25 chance of 3,000 alongside a 0.75 chance of 0. In this case, the implication is that the uncertainty is resolved in a single stage: perhaps a 20-sided die is rolled, and if a number from 1 to 4 comes up C pays 4,000 (and 0 otherwise), whereas D pays 3,000 if the number is anything in the range 1 to 5.

However, there is another way of presenting this choice which EU would regard as amounting to exactly the same thing, but which PT suggests people are likely to treat differently. Suppose that the uncertainty is resolved in two stages, as follows. In the first stage, there is a 0.75 chance of being 'knocked out' and getting 0, and there is a 0.25 chance of getting through to the second stage – at which point the choice is between, on the one hand, a 0.8 chance of 4,000 and, on the other hand, the certainty of 3,000. The logic of EU entails that the two stages can be 'reduced' to a single stage by multiplying through the probabilities: a 0.25 chance of getting through and facing a 0.8 chance of 4,000 can thus be reduced to a 0.2 chance of 4,000, as offered by prospect C; and a 0.25 chance of getting through to receive the certainty of 3,000 is regarded as just the same as a direct 0.25 chance of 3,000, as offered by prospect D.

Put another way, a 0.25 chance of what was prospect A in the common ratio example is equivalent to C, and a 0.25 chance of prospect B is regarded as the same as D. Yet the evidence of what K&T called the 'isolation effect' shows that people do not process the two-stage game in the way presumed by EU. When faced with such a two-stage problem and told that they have to make a commitment ahead of the first stage, most individuals appear to disregard (or isolate) the common first stage, focus on the alternatives that are contingent on getting through to the second stage, and then make much the same choices as they do when presented with the simple one-stage choice between A and B. In other words, when asked to commit ahead of this two-stage resolution of uncertainty, there is a much stronger tendency to pick the safer option than when presented with the one-stage choice between C and D where the calculus of probability 'reduction' has already been applied.

The fourth regularity that played a significant role in the formulation of PT was the 'reflection effect'. Essentially, this refers to the observations that changing payoffs from gains to losses (relative to the status quo) tended to reverse individuals' choices. Thus if A and B above were transformed into A′ and B′ such that A′ offered a 0.8 probability of *losing* 4,000 (and a 0.2 probability of losing nothing) while B′ entailed the certainty of a 3,000 loss, the modal preference for B over A would often be 'reflected' into a modal preference for A′ over B′. Thus, what appears as a predominant pattern of risk aversion in the choice between prospects such as A and B which involve gains seems to transform into a predominant pattern of risk seeking when the non-zero payoffs are losses.

Combining the reflection effect with the isolation effect can produce striking 'framing' effects. For example, consider first a scenario where an individual is given a lump sum of 1,000 and then asked to choose between a further 500 for sure or else a risky prospect offering a 50–50 chance of either 0 or an extra 1,000. If the individual isolates the initial 1,000 and displays risk aversion towards the 50–50 gamble involving gains, she will end up with a sure 1,500 rather than a portfolio consisting of a 0.5 chance of a net 1,000 and a 0.5 chance of a net 2,000. But now consider a scenario framed somewhat differently. The individual is given a lump sum of 2,000 and then asked to choose between the certainty of losing 500 or else a 50–50 chance of either losing 1,000 or losing 0. If the individual again isolates the lump-sum but now displays risk seeking towards the 50–50 gamble involving losses, she will end up with exactly the opposite portfolio preference: that is, she will choose the portfolio consisting of a 0.5 chance of a net 1,000 and a 0.5 chance of a net 2,000 rather than 1,500 for sure. K&T presented evidence which showed that this was indeed a strong tendency among those who answered their hypothetical questions framed in these various ways.

The Aims and Structure of Prospect Theory

PT can be seen as offering a descriptive (rather than a prescriptive/normative) model of a particular area of decision making. K&T were careful to specify the domain to which their model applied: it was a theory of *choice* over pairs of prospects each involving *no more than two non-zero payoffs* where the *objective probabilities were given* to decision makers. As formulated in the 1979 paper, the theory did *not* apply to valuation tasks (for example, tasks that asked people how much they would pay or accept in exchange for some risky prospect), nor to prospects involving larger numbers of possible payoffs, nor to cases where there was ambiguity about the likelihood of different events occurring (although in their concluding remarks K&T expressed some optimism that the model could be extended to accommodate the latter two features, while relevant valuations might be inferred via some iterative procedure involving a series of choices between a prospect and different sure sums). Most importantly, because it set out to provide an account of *actual* behaviour rather than a prescription for how decision makers *ought* 'rationally' to behave, PT allowed the possibility of patterns of behaviour that decision makers might wish to modify if they ever became aware of the 'inconsistencies' involved (although, in the absence of opportunities for such realization, the 'anomalies' implied by PT could be expected to occur and persist).

To *some* extent, the elimination of some potentially undesirable possible implications of PT was handled by dividing the modelling of people's decision processes into two phases: first, the editing phase, which involved simplifying prospects and screening out transparent transgressions of reasonable behaviour; and then the evaluation phase, in which the preferred alternative was identified.

The editing phase prepared the ground for the evaluation phase in various intuitively appealing ways. It involved the detection of *transparent* dominance and the discarding/rejection of dominated alternatives in such cases (while allowing the possibility that dominance might be violated if more complicated ways of presenting the prospects obscured the dominance relationship). There was also scope for some *simplification* of prospects (for example, *rounding* of payoffs and/or probabilities). In cases where there were transparently common and/or riskless components, these were liable to be *segregated* and/or *cancelled*. It was also supposed that, when a prospect offered the same payoff contingent on different events with separately expressed probabilities, those probabilities would be added together. For example, suppose a prospect offered a payoff of 100 if a card drawn at random from a standard pack of playing cards turned out to be a spade, and offered the same payoff if the card turned out to be a heart: then the probabilities of these two events – each 0.25 – would be *combined* to give an overall 0.5 chance of receiving 100. Finally, all payoffs were *coded* into gains or losses relative to some reference point – this latter normally being the status quo, although in some circumstances it might be otherwise (as discussed in the penultimate subsection of the 1979 paper).

The evaluation phase involved the interaction of two components: the *value* function, and the *decision weight* function.

A careful reading of the 1979 exposition makes it clear that the subjective value associated with a particular payoff should, strictly speaking, be expected to be a function of *two* factors: the asset position that constitutes the individual's reference point, and the positive or negative change from that point represented by the payoff in question. However, K&T argued that, over quite broad ranges of initial asset positions and for many practical purposes, it is sufficient to focus just on one argument, namely, the size of the gain or loss entailed by any particular payoff.

Drawing on existing evidence, including a substantial body of work from the realm of psychophysics, K&T argued that such a value function is characterized by two key characteristics.

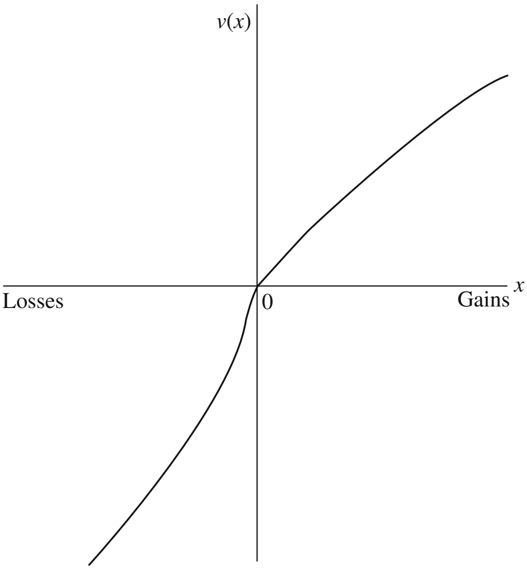
First, the marginal value of both gains and losses is presumed to diminish as the magnitudes increase. Thus the difference between a gain (or loss) of 100 and a gain (or loss) of 150 registers more strongly than the difference between a gain (loss) of 1,100 and a gain (loss) of 1,150. Such diminishing sensitivity means that the gradient of the value function becomes progressively less steep as payoffs are located further from the reference point. Denoting the value of any monetary payoff x by $v(x)$, diminishing sensitivity in the domain of gains can be more formally represented as

$$\begin{aligned} v(x+a)-v(x) > v(x+a+k) \\ - v(x+k) \text{ for all } x, a, k \\ > 0; \end{aligned}$$

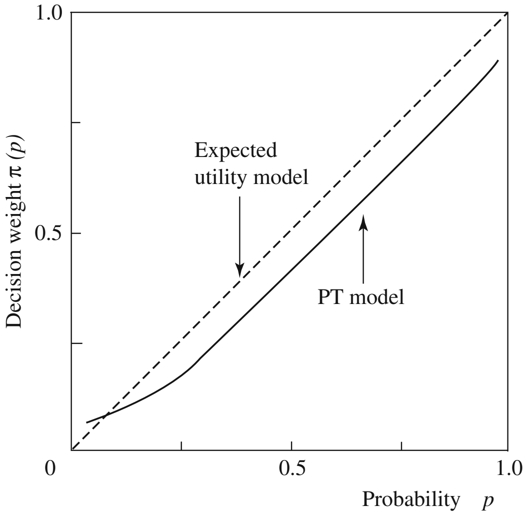
in the domain of losses, it entails

$$\begin{aligned} v(-x)-v(-x-a) > v(-x-k) \\ - v(-x-a-k). \end{aligned}$$

Second, the marginal value of losses is modelled as being greater than the marginal value of gains of the same magnitude: that is, for all x, the gradient of the function is steeper at $-x$ than at x. More formally, $v'(-x) > v'(x)$ wherever the derivative of x exists. In conjunction with the first characteristic, this implies a value function as



Prospect Theory, Fig. 1 Prospect theory's value function



Prospect Theory, Fig. 2 Prospect theory's decision weighting function

shown in Fig. 1: that is, concave in the domain of gains, convex and steeper in the domain of losses, and kinked at the (0) reference point.

Thus one part of the evaluation of any prospect involves converting money payoffs to their values via the $v(\cdot)$ function as specified above. The other part of the evaluation requires decision weights to be attached to the various values. These decision weights are not the objective probabilities, nor even degrees of belief of the kind that are conventionally supposed to constitute subjective probabilities. In the context of the 1979 exposition, they represent a psychological transformation, or modification, of the objectively given probabilities, with the weighting function being denoted by $\pi(\cdot)$.

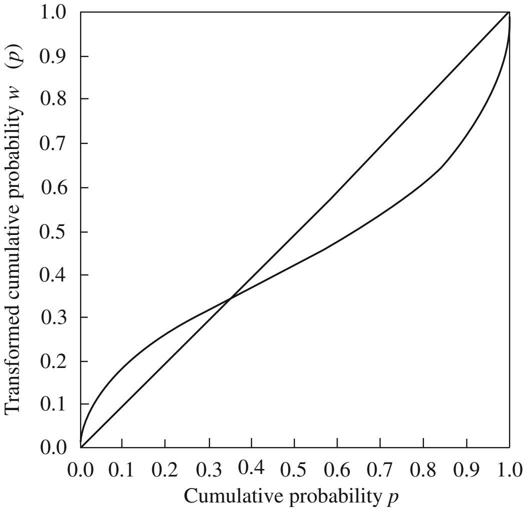
The key assumptions about $\pi(\cdot)$ are as follows. First, the weight attached to a zero probability event is 0, and the weight attached to a certainty is 1: that is, $\pi(0) = 0$ and $\pi(1) = 1$. Second, for low probability events, $\pi(p) > p$; but for higher probability events, $\pi(p) < p$; the 'crossover point', where $\pi(p) = p$, may vary from one individual to another, but is often depicted as being somewhere in the region of $p = 0.15$. Third, it is generally supposed that $\pi(p) + \pi(1 - p) < 1$: this property, labelled *subcertainty*, conveys the idea that complementary intermediate probabilities are jointly disadvantaged relative to certainty.

Taken together, the above assumptions are consistent with a decision weighting function of the kind depicted in Fig. 2. Over most of its range, the fact that $\pi(\cdot)$ is flatter than the 45 line suggests that the evaluation of a prospect is less sensitive to changes in the probability of its non-zero payoff (s) than would be the case under EU where the utilities of payoffs are weighted in exact proportion to their respective probabilities of occurring. It also has the implication that for any given ratio of probabilities, the ratio tends to get closer to 1 as the magnitudes of the probabilities fall: more formally, $\pi(pq)/\pi(p) \leq \pi(pqr)/\pi(pr)$ for all p , q , $r < 1$.

However, such a formulation also has the property that it entails abrupt changes/ jumps in the vicinities of $p = 0$ and $p = 1$. It might be said that the function is not 'well-behaved' – or indeed, not defined – in those regions, where there are 'quantal effects'. And this allows at least one pattern of behaviour that many decision theorists would find normatively undesirable/unacceptable, as follows. Consider a case where prospect C offers the certainty of some gain x, while A offers x with probability p and $x + a$ with probability $1 - p$, where a is some (small) positive amount. Evaluating each prospect separately,

$$\begin{aligned} v(\mathrm{C}) &= v(x)\,\text{and}\,v(\mathrm{A}) \\ &= \pi(p)v(x) + \pi(1-p)v(x+a). \end{aligned}$$

However, because $\pi(p) + \pi(1 - p)$ is liable to sum to less than 1, v(A) may be less than v(C), even though A dominates C. In a direct choice between the two, PT supposes that this dominance (if transparent) will be detected as part of the editing process, so that A will be chosen. But it should be possible to construct some other prospect B which neither dominates nor is dominated by either A or C but whose value lies between the two, so that $v(\mathrm{C}) > v(\mathrm{B}) > v(\mathrm{A})$. Hence in separate pairwise choices, C will be preferred to B and B will be preferred to A, while A will be chosen over C on the basis of transparent dominance, thereby giving a violation of transitivity.



Prospect Theory, Fig. 3 CPT's cumulative probability transformation function

More Recent Developments in Theory and Evidence

Because PT *does* allow such violations of principles that many decision theorists regard as normatively compelling, various modifications have been proposed to 'fix' this supposed defect: in particular, a method of deriving decision weights which ensured that they summed to 1 and disallowed any violations of stochastic dominance or transitivity was proposed by Quiggin (1982) and was subsequently incorporated into a revised and extended form of PT known as cumulative prospect theory (CPT) (see Tversky and Kahneman 1992).

The essence of Quiggin's proposal involved ranking the possible outcomes $x_1 \ldots x_n$ offered by a prospect according to their values and then assigning weights to each of the cumulative probabilities that the prospect pays at least xi, for all $i = 1 \ldots n$. (Hence this kind of model came to be labelled as 'rank-dependent'.) The function used to transform cumulative probabilities – call it $w(\cdot)$ to distinguish it from the $\pi(\cdot)$ discussed above – is fully defined in [0, 1] space, with $w(0) = 0$ and $w(1) = 1$. Like $\pi(\cdot)$, it is usually supposed to have an increasing inverse-S shape (although by contrast with $\pi(\cdot)$, the 'crossover point' in CPT is more often regarded as lying in the 0.3–0.4 region – see Fig. 3.

As a consequence of being steeper in the vicinities of 0 and 1, and less steep across the intermediate range, this form of $w(\cdot)$ gives greater weight to extreme than to intermediate outcomes. Although it may be psychologically implausible that most individuals transform probabilities *strictly* according to the rather cumbersome procedure specified by CPT and other rank-dependent models, the approach captures the general intuition that extreme outcomes may attract more attention and receive relatively more weight in decisions. And it appeals to those theorists who are inclined towards models that respect what are perceived to be 'fundamental' requirements of rationality such as transitivity, while also having the advantage of being applicable to prospects involving probability distributions over any number and range of outcomes.

However, the spirit of PT was to provide a *descriptive* model of risky choice, so that violations of transitivity of the kind outlined earlier were an implication of the model; and, to the extent that they occur in practice, PT can claim to be descriptively successful. And indeed there is evidence of such violations (see Starmer 1999).

On the other hand, as K&T acknowledge, there are limitations to the scope of PT. As discussed above, the domain of the theory was very specific and excluded a number of tasks that are of economic significance (such as the formulation of certainty equivalence values). Moreover, certain assumptions made in the model are open to question. For example, the phenomenon of 'event-splitting' (see Humphrey 1995) suggests that people may only imperfectly add the probabilities of the same payoff under different 'states of the world', contrary to the supposed process of combination in the editing phase. There may also be questions about just how transparent dominance needs to be before it is detected in the editing phase. And some researchers – see, for example, Birnbaum (2006) – have amassed evidence of patterns of choice which appear to run counter to the claims of prospect theories to provide a satisfactory description even of the behaviour which should lie within their domain.

All that having been said, there can be no doubt whatsoever of the success of PT in focusing the attention of decision theorists on patterns of behaviour that do not conform with the conventional (and still predominant) wisdom of EU, and in stimulating a very substantial body of experimental, empirical and theoretical work exploring behaviour outside of the strictures of standard economic 'rationality'.

See Also

- ▶ Expected Utility Hypothesis
- ▶ Experimental Economics
- ▶ Kahneman, Daniel (Born 1934)
- ▶ Non-expected Utility Theory

Bibliography

Allais, M. 1953. Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école américaine. *Econometrica* 21: 503–546.

Birnbaum, M. 2006. Evidence against prospect theories in gambles with positive, negative and mixed consequences. *Journal of Economic Psychology* 27: 737–761.

Humphrey, S. 1995. Risk aversion or event-splitting effects? More evidence under risk and uncertainty. *Journal of Risk and Uncertainty* 11: 263–274.

Kahneman, D., and A. Tversky. 1979. Prospect theory: An analysis of decision under risk. *Econometrica* 47: 263–291.

Kim, E., A. Morse, and L. Zingales. 2006. What has mattered to economics since 1970. *Journal of Economic Perspectives* 20(4): 189–202.

Quiggin, J. 1982. A theory of anticipated utility. *Journal of Economic Behavior and Organization* 3: 323–343.

Starmer, C. 1999. Cycling with rules of thumb: An experimental test for a new form of non-transitive behaviour. *Theory and Decision* 46: 139–157.

Starmer, C. 2000. Developments in non-expected utility theory: The hunt for a descriptive theory of choice under risk. *Journal of Economic Literature* 38: 332–382.

Tversky, A., and D. Kahneman. 1992. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty* 5: 297–323.

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

Protoindustrialization

Sheilagh Ogilvie

Abstract

'Proto-industrialization' is the name given to the massive expansion of export-oriented handicrafts which took place in many parts of Europe between the 16th and the 19th centuries. An influential theory holds that these proto-industries generated the capital, labour, entrepreneurship, agricultural commercialization, and consumer demand needed for factory industrialization. Protoindustrialization, it is argued, also transformed traditional economic mentalities and institutions. However, deeper empirical study has cast doubt on most of these claims. Theories of protoindustrialization have stimulated much excellent research, but do not explain the significant economic growth, demographic change, and institutional transformation that occurred in Europe before the Industrial Revolution.

Keywords

Agriculture; Fertility; Industrial revolution; Industrialization; Nuptiality; Population growth; Protoindustrialization; Putting-out system; Surplus labour; Urbanization; Workshop system

JEL Classifications

N50; N01; N33; N53; N63; N73; N93

'Protoindustrialization' is the name given to the massive expansion of export-oriented handicrafts which took place in many parts of Europe between the 16th and the 19th centuries.

Often, although not always, such proto-industries arose in the countryside where they were practised alongside agriculture; usually, they expanded without adopting new techniques or centralizing production into factories. This growth of pre-factory industry in early modern Europe has long been a subject of specialized study. But in the 1970s it began to attract much wider interest, when several influential works christened it 'protoindustrialization' and argued that it was a major cause of industrialization and capitalism.

Protoindustrialization as the First Stage of Industrialization

The term 'protoindustrialization' was invented by Franklin Mendels, who first used it in his 1969 dissertation on the Flemish linen industry (published in 1981) and popularized it in a now famous article based on that research (Mendels 1972). Mendels claimed that protoindustrialization was the first phase of industrialization. In the 18th century, seasonally underemployed European country-dwellers moved massively into cottage crafts, exporting their wares beyond the immediate region. This, Mendels argued, broke down traditional urban institutions such as guilds that had previously limited industrial growth. Mendels contended that it also weakened rural institutions such as inheritance systems, communes, and manorial systems that had traditionally calibrated population growth to economic resources. Mendels claimed that this made nuptiality (and thus fertility) ratchet upwards: proto-industrial upswings saw more marriages, but downswings did not see fewer. High protoindustrial fertility fuelled rapid population growth, Mendels argued, in turn causing further industrial expansion. This self-sustaining proto-industrial spiral, according to Mendels, generated the capital, labour, entrepreneurship, agricultural commercialization, and consumer demand needed for factory industrialization.

Protoindustrialization and Proletarianization

Mendels's arguments were initially widely adopted, giving rise to several schools of protoindustrial theory. One emanated from David Levine, whose study of two villages in 19th-century Leicestershire appeared to confirm that proto-industry led to population growth (Levine 1977). For Levine, proto-industry was important mainly because he believed it broke down rural social structure and land ownership, creating a large group of landless people who had to work for wages. This broader process of 'proletarianization' was, Levine argued, crucial for capitalism and industrialization.

Protoindustrialization and Surplus Labour

A third view of protoindustrialization was put forward by Joel Mokyr (1976), who rejected almost all the arguments advanced by Mendels and Levine but argued that proto-industries provided the cheap 'surplus' labour to fuel a 'dualistic' growth of the European economy as modelled for modern less developed countries (LDCs) by Lewis (1954) and Fei and Ranis (1964). The key empirical problem for the Lewis–Fei–Ranis model was whether 'surplus' labour existed and where it came from. Mokyr argued that in pre-industrial Europe surplus labour came from protoindustry, creating a flat labour supply curve and hence very low wages for early factory industry. This 'dualistic labour surplus' view of protoindustrialization has hardly been pursued empirically, but is important because of its links with development economics and with Jan De Vries's influential theory of European urbanization (De Vries 1984).

Protoindustrialization and the Transition to Capitalism

The protoindustrialization debate was intensified by the publication of a massive book by Kriedte et al. (1977 (German original), 1981 (English translation)). Combining Mendels's and Levine's findings with the voluminous earlier literature on cottage industries, these scholars turned the theory of protoindustrialization into a general model of European economic transformation between the medieval and modern periods.

For them, protoindustrialization was the 'second phase' of this transformation process. The first phase, they claimed, was a loosening of feudalism caused by commutation of feudal burdens from labour or grain dues into money rents, polarizing the rural population into two classes: well-off peasants with enough land to live solely from farming, and land-poor or landless strata who had to seek work outside agriculture. The second phase, in their view, was the 16th-century growth in supra-regional and international trade, creating a growing demand for manufactures which the new rural proletariat could satisfy more cheaply than guild-regulated urban craftsmen. So protoindustries arose in the countryside.

These scholars proposed a stage theory according to which rural protoindustries then gradually transformed industrial organization. The first stage, they claimed, was the *Kaufsystem* (artisanal or workshop system), in which rural producers retained autonomy over production and selling. The second stage, the argument continues, was the *Verlagssystem* (putting-out system), in which merchants bought raw materials, 'put them out' to the rural producers who processed them in return for a wage, and then collected the output for transfer either to the finishing stages of production or to the final consumer market. This ultimately led to a third stage, it was claimed: the concentration and mechanization of production in centralized, mechanized factories.

Extensions to the Theories of Protoindustrialization

By 1977 at the latest, therefore, protoindustrialization had generated a family of different theories, based on differing definitions of protoindustry and differing explanations of economic development. Almost all they had in common was to emphasize the significance of European economic and demographic growth before factory industrialization, and to ascribe such growth to changes in a certain economic sector – export-oriented cottage industry. Over the following decades, these various branches of protoindustrialization theory stimulated a huge outpouring of research into pre-industrial manufacturing, not just in Europe but also in the non-European world, including modern LDCs.

By 1982 protoindustrialization had become such an influential concept that Franklin Mendels and Pierre Deyon were invited to convene one of the three main sessions of the Eighth International Economic History Congress in Budapest, with protoindustrialization as their theme. They pre-circulated a set of hypotheses, 48 researchers contributed papers (Deyon and Mendels 1982), and Mendels summarized the session with a report, a revised definition, and a set of hypotheses for subsequent debate (Mendels 1982).

This new 1982 definition of protoindustrialization stressed five key characteristics. First, protoindustrialization occurred not nationally or internationally, but *regionally*: 'within a small radius around a regional capital'. Second, protoindustries must be distinguished from traditional crafts: they produced not for local or regional consumption, but for sale to *export* markets outside the region. Third, protoindustry was mainly *rural* and *part-time* – only in its final or extreme phase did it involve full-time industrial employment. Fourth, protoindustrialization arose symbiotically with *agricultural commercialization*. Finally, protoindustrialization was 'dynamic': it was defined as a *growth* over time in the industrial employment of rural workers.

Deyon and Mendels also proposed four central hypotheses about the *effects* of protoindustrialization. First, protoindustry led to population growth and land fragmentation because it broke down traditional demographic regulation by communes, landlords and inheritance systems. Second, protoindustrial profits created the capital for factory industrialization. Third, protoindustry trained merchants and workers in the skills needed for factory industrialization. Finally, protoindustrialization caused agriculture to commercialize, thereby feeding urbanization and industrialization. Through these four mechanisms, proto-industry led to factory industry – although the authors admitted that sometimes it led to de-industrialization instead (Mendels 1982).

Criticisms of the Theories of Protoindustrialization

Somewhat more slowly than they attracted support, the theories of protoindustrialization also began to draw criticism.

For one thing, the precise size and structure of the unit that qualified as a protoindustrial *region* was unclear. Protoindustries could and often did extend beyond the radius around a single market town, or alternatively were sometimes found in only one or two communities in such a radius. One pragmatic solution was to define the region as simply the area within which a certain proto-industry was practised. But this seemed to leach the concept of the region of much of its analytical content. Second, there was no agreement about how large a proportion of the regional labour force must have been employed in protoindustry, or how fast or sustained the growth of this labour force must have been, in order to qualify as 'protoindustrialization' (Ogilvie and Cerman 1996).

There was also confusion about the precise importance of *export markets* for protoindustrialization. First, why were export markets uniquely important? Second, what proportion of production had to be exported in order for any given industry to qualify as a proto-industry instead of a craft? Third, how distant did final markets have to be to qualify as 'supra-regional' rather than 'local'? The demarcation between local crafts and export-oriented proto-industries was thus very unclear and its analytical importance remained obscure.

The *neglect of other forms of industry* was another weak point. The theories of protoindustrialization concentrated solely on one sort of pre-industrial manufacturing: cottage industry. But what justified this emphasis? Did manufacturing really develop just because of this single sort of industry, which was often technologically very primitive? What about highly skilled and technologically innovative crafts, export-oriented urban industries, or centralized manufactories? Mainstream historians of pre-industrial manufacturing argued that all these branches of the secondary sector should be included in any analysis of industrialization before the Industrial Revolution (Schremmer 1981; Coleman 1983; Mager 1993). Others argued that large *urban* export industries, and those involving centralized production units, should also be included under the rubric of protoindustrialization (Cerman 1993).

The *neglect of industrial technology and physical geography* was also criticized. Mendels referred in passing to industrial production functions and transportation costs, but neither he nor other proponents of the theory explored these factors further. Critics argued that any coherent view of protoindustry must consider the technical requirements of different branches of industry and the geographical and physical characteristics of the region (Mager 1993). Others urged that protoindustry, like any economic activity, be analysed in terms of 'opportunity costs', and pointed out that this would imply taking into account a whole array of technological, geographical, and institutional variables (Ogilvie 1993; 1997).

The theories adopted strong assumptions about the '*traditional societies*' transformed by proto-industry, and these assumptions began to be questioned (Coleman 1983; Houston and Snell 1984; Schremmer 1981; Ogilvie and Cerman 1996; Ogilvie 1997). Protoindustrialization theorists had uncritically accepted the theories of

Alexander Chayanov, who regarded peasants as unable and unwilling to calculate costs, seek profits, use money, or transact in markets (Chayanov 1966). But was this really true of the early modern European rural population? The subsistence-orientation assumed for rural domestic workers was not confirmed by empirical studies, and was inconsistent with the fact that proto-industrial producers often became traders, middlemen, putters-out and even manufactory-operators. The demographic decisions and productive choices of protoindustrial workers, rather than being governed by 'traditional mentalities', began to look highly rational (Ogilvie 1997).

The *demographic* predictions of the theories were widely falsified as empirical studies proliferated. It emerged that pre-industrial demographic behaviour was influenced by such a wide array of variables that proto-industry could have highly divergent effects on nuptiality, fertility, mortality and migration in different European societies. Case studies showed that not all protoindustrial regions had greater population density, faster demographic growth, lower ages of marriage, higher fertility rates, larger households, or a breakdown in the family and gender division of labour – all of which had been postulated in the original theories. Furthermore, many – even all – of these demographic changes could also be observed in some primarily agricultural regions (Schremmer 1981; Coleman 1983; Houston and Snell 1984; Ogilvie and Cerman 1996; Ogilvie 1997).

The relationship between *commercial agriculture* and protoindustry was also disputed. Protoindustries arose alongside many different kinds of agriculture, including subsistence cultivation, market farming, and even large feudal domains worked by serf labour. Protoindustries derived food and raw materials not just from commercial agriculture but from local cultivation by proto-industrial workers themselves. Simultaneous employment in proto-industry and agriculture was common but not universal in proto-industrial regions. While traditional agrarian institutions and rural social structure broke down in some protoindustrial regions, in others they survived unaltered for centuries (Houston and Snell 1984; Ogilvie and Cerman 1996; Ogilvie 1997).

The role of *social and political institutions* in theories of protoindustrialization has also been critically revised (Ogilvie 1993; Ogilvie and Cerman 1996; Ogilvie 1997; Ogilvie 2004). The original theorists assumed that protoindustrialization both required and furthered the replacement of 'traditional' social institutions with markets. But deeper research has shown that urban privileges, craft guilds, monopolistic merchant companies, village communities and manorial institutions remained important in many European protoindustries, and crucially influenced economic, demographic and social change in proto-industrial regions.

A final major criticism questioned the role of proto-industry in causing *factory industrialization*. Each of the mechanisms by which protoindustrialization is supposed to have led to industrialization has been subject to sceptical re-evaluation. Research shows that the demographic effects of protoindustrialization were extremely various, as was its impact on the fragmentation of landholdings. Protoindustry appears to have been only one of many sources of capital invested in the early factories, and in many cases proto-industrial profits flowed into agriculture, landholding or socio-political investments. Proto-industry was also only one of many sources of entrepreneurial skills for industrialization, and sometimes did not encourage entrepreneurship at all. There is little evidence that it was proto-industry that led to commercial agriculture rather than that agricultural surpluses made possible both proto-industrial regions and urbanization. It is now widely acknowledged by both the theorists and their critics that proto-industry often led not to factories but to de-industrialization and a return to agriculture. The critics argue that this finding denudes the theory of most of its empirical content (Coleman 1983; Houston and Snell 1984; Clarkson 1985; Ogilvie and Cerman 1996). Although, therefore, the theory of protoindustrialization has stimulated much excellent research, it does not explain the significant economic growth, demographic change, and socio-institutional transformation that indisputably occurred in Europe well before the industrial revolution.

See Also

- ▶ Agriculture and Economic Development
- ▶ Capitalism
- ▶ Development Economics
- ▶ Dual Economies
- ▶ Economic Demography
- ▶ Economic History
- ▶ Growth and Institutions
- ▶ Historical Demography
- ▶ Industrial Revolution
- ▶ Labour Surplus Economies
- ▶ Lewis, W. Arthur (1915–1991)
- ▶ Medieval Guilds
- ▶ Peasants

Bibliography

Cerman, M. 1993. Protoindustrialization in an urban environment: Vienna, 1750–1857. *Continuity and Change* 8: 281–320.

Chayanov, A. 1966. *The theory of peasant economy*, ed. D. Thorner, B. Kerblay, and R. E. F. Smith. Homewood: Richard A. Irwin.

Clarkson, L. 1985. *Proto-industrialization: The first phase of industrialization?* Houndmills: Macmillan.

Coleman, D. 1983. Protoindustrialization: A concept too many. *Economic History Review* 36(2nd series): 435–448.

De Vries, J. 1984. *European urbanization 1500–1800*. Cambridge, MA: Harvard University Press.

Deyon, P., and F. Mendels (eds.). 1982. *La protoindustrialisation: Théorie et réalité*. Lille: Université de Lille.

Fei, J., and G. Ranis. 1964. *Development of the labour surplus economy: Theory and policy*. Homewood: Richard A. Irwin.

Houston, R., and K. Snell. 1984. Proto-industrialisation? Cottage industry, social change and industrial revolution. *Historical Journal* 27: 473–492.

Kriedte, P., H. Medick, and J. Schlumbohm. 1977. *Industrialisierung vor der Industrialisierung. Gewerbliche Warenproduktion auf dem Land in der Formationsperiode des Kapitalismus*. Göttingen: Vandenhoeck & Ruprecht.

Kriedte, P., H. Medick, and J. Schlumbohm. 1981. *Industrialization before industrialization: Rural industry in the genesis of capitalism*. Cambridge: Cambridge University Press.

Levine, D. 1977. *Family formation in an age of nascent capitalism*. New York/London: Academic.

Lewis, W. 1954. Economic development with unlimited supplies of labour. *Manchester School of Economic and Social Studies* 22: 139–191.

Mager, W. 1993. Protoindustrialization and proto-industry: The uses and drawbacks of two concepts. *Continuity and Change* 8: 181–216.

Mendels, F. 1972. Protoindustrialization: The first phase of the industrialization process. *Journal of Economic History* 32: 241–261.

Mendels, F. 1981. *Industrialization and population pressure in eighteenth-century flanders*. New York: Arno Press.

Mendels, F. 1982. Protoindustrialization: Theory and reality. General report. In *Eighth international economic history congress, Budapest 1982: 'A' themes*, ed. F. Mendels and P. Deyon. Budapest: Akademiai Kiado.

Mokyr, J. 1976. Growing-up and the industrial revolution in Europe. *Explorations in Economic History* 13: 371–396.

Ogilvie, S. 1993. Protoindustrialization in Europe. *Continuity and Change* 8: 159–179.

Ogilvie, S. 1996. Social institutions and protoindustrialization. In *European protoindustrialization*, ed. S. Ogilvie and M. Cerman. Cambridge: Cambridge University Press.

Ogilvie, S. 1997. *State corporatism and proto-industry: The Württemberg Black Forest, 1580–1797*. Cambridge: Cambridge University Press.

Ogilvie, S. 2004. Guilds, efficiency and social capital: Evidence from German protoindustry. *Economic History Review* 57: 286–333.

Ogilvie, S., and M. Cerman. 1996. Protoindustrialization, economic development and social change in early modern Europe. In *European proto-industrialization*, ed. S. Ogilvie and M. Cerman. Cambridge: Cambridge University Press.

Schremmer, E. 1981. Proto-industrialisation: A step towards industrialisation? *Journal of European Economic History* 10: 653–670.

Proudhon, Pierre Joseph (1809–1865)

H. Bartoli

Proudhon was born in Besançon, France, into a very humble family. Despite a scholarship, poverty forced him to interrupt his exceptionally brilliant studies. He became, in turn, a printer, print shop foreman, scholarship student at the Besançon Academy, owner of a small print shop, and managing clerk in a river transport company in Lyons. He then became a writer and journalist, following this profession through incessant material difficulties, political trials, election to parliament, prison and exile. On his death he left a vast body of work, in which he tackled at the same time problems of philosophy, ethics, sociology and economics. He can equally well be seen as one of the founders of sociology, the father of anarchism, one of the inspirational forces behind cooperativism and mutualism, one of the sources of syndicalist thinking, 'the boldest thinker of French socialism' (Marx), a pioneer of federalism and regionalism, or one of the apostles of mass education.

From his study of history and observation of the world, Proudhon derived a 'serial dialectic'. Everything in the world is 'serial' – i.e. is differentiated, divided, graduated and graded, but also coordinated, articulated, grouped; everything is multiple, everything is synthesis. The 'series' is a 'whole composed of elements arranged according to a certain reason or law'. 'Serial dialectic' is a 'law' of progression and organization, a general process of growth common to matter and spirit, to man and society. An antinomic dialectic, it unfolds as a chain of antinomic pairs whose opposition is the source of all movement and cannot be resolved into synthesis. Such a dialectic of tension, or of the 'balancing of opposites', is thus fundamentally opposed to the Marxian dialectic of synthesis.

To struggle to be, to unite to be, are the two poles of the vital dialectic of every person and every society. Work is the condition for survival, the constituent 'organic law', the) 'generative fact', the 'shaping force' of society. Antagonism and solidarity are no more than 'functional laws'.

All labour implies at the same time differentiation and association. In working society one does not find 'workers' but a single worker diversified to infinity. The 'fundamental' law of labour is the law of division. There is a further law connected with this – that of 'collective force' as expressed in the) 'collective surplus' generated by association, the collective product being the result not of the addition of individual efforts, but of their multiplication when they are brought together in association.

Labour, for Proudhon, is the 'field of observation' of political economy, which studies the division of labour and its series (organization, collective force), the distribution of the instruments of labour (right and mutuality), and the efficiency of labour and its results (value and economic accounting). To organize labour is to demarcate functions, and then to group them according to the laws of labour. The division of labour is the law of function; every individual worker is therefore necessarily an integral member of the enterprise, fulfilling an economic function. Starting from these individual functions, through a kind of integration, one can organize the entire society.

Labour is the real measure of exchange value, the only standard by which different products can be compared. The substance common to wages, investments, capital and profits is that they are either objectified labour or accumulated labour. Supply and demand are simply '*éléments traducteurs*', (translating factors) constantly disrupted by monopolies, fraud, speculation etc., and do not allow use value (utility) and exchange value (labour costs) to be objectively compared. For the 'law of proportionality of values' to be respected, a) 'constituted value', a synthesis of use value and exchange value, must be created; 'society's accounts' must also be drawn up, labour scientifically managed and the structure 'socialized'.

Proudhon regarded accounting as 'the whole of political economy'. An astonishing forerunner of many later theories, he drew a distinction between individual accounts and accounts relating to 'each type of value' ('chaque nature de valeurs') and combined them in a 'single account', a veritable set of national economic and social accounts. He hoped to establish a form of accounting by sector and by industry, prelude to a 'centralization of accounts', since 'all industries are bound together in one cluster by their mutual relations . . . all products act as ends and as means of each other'. He glimpsed the problem of variation in the 'proportion' of labour, and hence in input–output coefficients, and he connected this to technical progress. He believed that a kind of) 'higher mathematics' could help in developing 'social economics', but warned against any ill-considered use of mathematics in economics.

All the ills of mankind spring from 'mere accounting error'. The 'social balance' is inexact. The gratuitous appropriation of collective effort, the inequality of exchange, the law of escheat, all distort the economic accounts. Property is at the same time) 'right of exclusion and theft' and 'despotic power'. Competition, although a necessary stimulant, kills competition; it generates monopoly, which is necessary in that it consolidates the achievements of labour, but which corrupts economic life since it improperly appropriates to itself the profits of) 'collective force' and creates poverty.

Proudhon's historical labourism bears no relation to Marx's historical materialism. For Proudhon, social and economic facts are only the 'manifestations' and 'signs' of ideas. Economics is metaphysics in action, the implementation of the) 'eternal laws of reason'. Proudhon declared himself against capitalism, the exploitation of man by man; against statism, the government of man by man; against communism, 'the degradation of the personality in the name of society', and against Christianity, 'a system of personal degradation in the name of right'. All his works are a prodigious effort to lay bare the foundations, the elements and the method for a self-managed society free of all alienation. He foresaw and proposed the building of a 'scientific socialism'.

What makes society possible, for Proudhon, is the) 'opposition of powers', 'mutual counterbalance'. Society must be organized solely on the basis of contract. In industry, wage labour will be replaced by common, joint ownership by all those who play a part in production, while in agriculture individually owned farms will be integrated into communes or agricultural groups, and cooperatives will prevail in trade and commerce. Through the federation of 'business properties' (*propriétés d'entreprises*) and rural communes and the establishment of consumers' associations and a production/consumption union, a federative republic can be created, the government of which would be formed by successive delegations from 'natural', autonomous, self-managed groups. Within this industrial republic the equity of social relations will be assured by free credit and 'exchange vouchers' issued by an exchange bank and secured against products.

Proudhon had faith only in the 'proletarians' to bring into being this new social structure, but unlike Marx he saw the emancipation of the proletariat merely as a particular fact of world history 'which is in the process of taking place'.

Selected Works

1867–70. *Oeuvres complètes*, 26 vols. Paris: Lacroix; Brussels; Verboeckhoven et Cie. 2nd edn, Paris: Rivière, with previously unpublished notes and texts.

1953. *Textes choisis*, Ed. J. Lajugie. Paris: Dalloz.

1967. *Oeuvres choisies*, Ed. J. Bancal. Paris: Gallimard.

Bibliography

Bancal, J. 1970. *Proudhon, pluralisme et autogestion*. Paris: Aubier-Montaigne.

Bancal, J., et al. 1967. *L'actualité de Proudhon*. Brussels: Institut de Sociologie.

Bouglé, C. 1930. *Proudhon*. Paris: Alcan.

Diehl, K. 1888–96. *P.J. Proudhon, seine Lehre und sein Leben*, 3 vols. Jena: Fischer.

Gurvitch, G. 1965. *Proudhon*. Paris: Press Universitaires de France.

Haubtmann, P. 1980. *La philosophie sociale de P.J. Proudhon*. Grenoble: Presses Universitaires.

Haubtmann, P. 1981. *Proudhon, Marx et la pensée allemande*. Grenoble: Presses Universitaires.

Haubtmann, P. 1983. *P.J. Proudhon, sa vie et sa pensée*. Paris: Beachesne.

Marx, K. 1847. *The poverty of philosophy.* (A critique of Proudhon's *Philosophie de la misère*.). In *Collected works*, vol. VI. Moscow: Progress Publishers, 1976.

Woodcock, G. 1956. *P.J. Proudhon: A biography*. London: Routledge & Kegan Paul.

Pseudo-distribution

Peter Groenewegen

A term devised by Edwin Cannan (1893) to distinguish the analysis of 'wages per head, profits per cent and rent per acre' (pseudo-distribution) from what he called) 'distribution proper' or analysis of 'division of the whole produce between aggregate wages, aggregate profits and aggregate rents' (Cannan 1893, p. 267). Cannan's argument in support of this distinction had both historical and analytical aspects while he also argued (1905) that it conformed to the meaning of 'division' ordinarily assigned to the word) 'distribution', in which 'a change in distribution', for instance, is taken to mean 'a change in the proportions in which the total is divided'.

His analytical reason for the distinction was as follows. 'In the equation, produce = wages + profits + rents,. . . the question should be to determine what settles the relative magnitude of the three terms on the other side of the equation' (Cannan 1893, p. 180). This problem only makes sense when aggregate wages, rents and profits are considered, and these are clearly not identical with changes in wages, profit and rent as rates of renumeration. When analysis concentrates exclusively on relative factor prices or 'pseudo-distribution', some analytical insights may be lost. In his last book, Cannan (1929, pp. 301–2) explained why in 1905 he had defended the role of) 'distribution proper' in economic analysis to combat what by then had become the traditional theory of distribution. Integrating the determination of wage, profit and rent rates with the supply and demand theory of value was insufficient to answer the distributional questions in which people and policy makers were interested, since by itself this could not explain variations in income shares over time.

Cannan implied that his distinction was particularly appropriate for historical investigation, presumably the reason why it surfaced in his history of production and distribution theories. However, although Quesnay's use of the term in the *Tableau économique* suggests shares as the essential feature of the distribution problem, Turgot's use of the term and that by Smith concentrates on rate of renumeration aspects. Even Ricardo (1817, p. 5), who defined distribution – 'the principal problem in Political Economy' – in terms of dividing total produce among landlords, labourers and capitalists, analytically emphasized the importance of determining the rate of profit as against the profit share. However, irrespective of the importance of treating distribution by means of analysing relative income shares accruing to social classes, by the end of the century major concern with such issues had disappeared and distribution was analysed exclusively in terms of rates of reward.

Few economists appear to have followed Cannan's terminology. Dalton (1920, pp. 136–9) approvingly referred to it; Robbins (1932, pp. 64–5) explicitly rejected it. By the end of the 1930s, Stigler (1941, p. 3) could refer to a classical 'failure to develop a theory of the prices of productive services' as confirmation of the fact that 'in 1870 there was no theory of distribution'. Only some catholic approaches to distribution theory since then (Bronfenbrenner 1971; Pen 1971) have recognized Cannan's complaint about confining distribution analysis to factor pricing. In Bronfenbrenner's case (1971, p. 121) this is combined with the apology that his treatment of the topic maintains stress on 'pseudo-distribution as modified aggregately since Cannan's day' by the type of distribution theory pioneered by Keynes and Kalecki, though it may also be noted that from the 1920s distributional consequences in terms of relative shares were being drawn from Cobb–Douglas production functions.

See Also

► Cannan, Edwin (1861–1935)

Bibliography

Bronfenbrenner, M. 1971. *Income distribution theory.* London: Macmillan.

Cannan, E. 1893. *A history of the theories of production and distribution from 1776 to 1848.* Reprinted from 3rd edn. London: Staples, 1953.

Cannan, E. 1905. The division of income. *Quarterly Journal of Economics* 19: 341–369.

Cannan, E. 1929. *A review of economic theory.* London: P.S. King.

Dalton, H. 1920. *Some aspects of the inequality of incomes in modern communities.* London: Routledge, 1935.

Pen, J. 1971. *Income distribution.* Harmondsworth: Pelican Books, 1974.

Ricardo, D. 1817. *Principles of political economy and taxation*, ed. P. Sraffa. Cambridge: Cambridge University Press, 1951.

Robbins, L. 1932. *An essay on the nature and significance of economic science*. London: Macmillan, 1949.
Stigler, G.J. 1941. *Production and distribution theories: The formative period*. New York: Macmillan.

Psychological Games

Martin Dufwenberg

Abstract

Traditional game-theoretic models assume that utilities depend only on actions. This is not sufficient for describing the motivations and choices of decision makers who care about reciprocity, emotions, or social rewards. Psychological games allow utilities to depend directly on beliefs (about beliefs) in addition to which actions are chosen, and they can capture a wider range of motivations. This article contains several examples and it is indicated where research on psychological games is headed.

Keywords

Allais Paradox; Belief-Dependent Motivation; Commitment; Decision Theory; Emotions; Extensive Game Forms; Game Theory; Guilt Aversion; Psychological Forward Induction; Psychological Games; Reciprocity; Signalling; Trust; Von Neumann and Morgenstern

JEL Classifications

C9

Traditional game-theoretic models presume that utilities depend on actions. While this framework is quite general (it can, for example, accommodate profit- maximization, altruism, inequity aversion and Rawlsian maximin preferences) it is not rich enough to adequately describe several psychological or social aspects of motivation which depend directly on beliefs (about beliefs) in addition to which actions are chosen. The following example illustrates.

Karen feels guilty if she lets others down. When paying her landscaper (Jim), this influences her tipping. The more she believes Jim believes he will receive as a tip, the more she gives. More precisely, she gives just as much as she believes Jim believes he will get, in order to avoid the feelings of guilt that will plague her if she gives less.

Beyond depicting something arguably realistic, the example illustrates in the simplest possible way how one may have to transcend traditional game theory to model a belief-dependent motivation. Consider a standard game form where Karen chooses a tip t such that $0 \leq t \leq w$, where w is the number of dollars in her wallet, and where the landscaper has no choice (his strategy set is modelled as a singleton $\{x\}$). Karen's choice of tip thus pins down a strategy profile (t,x). In traditional game theory, payoffs are defined on strategy profiles (or on endnodes induced by strategy profiles), so Karen's best choice (or choices) would be independent of her belief about Jim's belief about her choice of tip. This runs counter to the example.

Gilboa and Schmeidler (1988) and Geanakoplos, Pearce and Stacchetti (Geanakoplos et al. 1989) present several examples that illustrate the inadequacy of traditional methods of representing preferences that reflect various forms of belief-dependent motivation. Geanakoplos, Pearce and Stacchetti develop a new analytical framework, in which the centrepiece is the notion of a *psychological game,* which may be seen as a generalization of a traditional game and which can model some of the desired effects. A psychological game differs from a traditional game in that utilities are defined on beliefs (about actions and beliefs), as well as on which actions are chosen. (The term 'game with belief-dependent motivation' would be more descriptive than the term 'psychological game', but I stick with the latter, which has become established.)

Reciprocity

The best-known example of a psychological games-based application is Rabin's (1993) highly influential model of reciprocity, according to which players wish to act kindly (unkindly) in response to kind (unkind) actions. The key notion of kindness depends on beliefs in such a way that reciprocal motivation can be described only by using psychological games. To see why, suppose that I jump out in front of your car, blocking your way, so that you can't cross a bridge and therefore arrive late to an important meeting. Am I kind? Clearly one cannot say without knowing what my beliefs are. If I believe the bridge is as sturdy as bridges usually are and I am just goofing around, then I am unkind. However, if I believe the bridge is about to collapse, then I am kind. Arguably, I would be kind even if I mistook a sturdy bridge for a dangerous one. So should you be kind or unkind in return? The answer depends on your beliefs about my kindness, and hence on your beliefs about my beliefs. It takes a psychological game to model that. (The example given here is similar in spirit to another example given in Rabin 1998, p. 23. Rabin's model is normal-form based. See Dufwenberg and Kirchsteiger 2004, for an extension to extensive game forms. See Fehr and Gächter 2000, and Sobel 2005, for general discussions of why reciprocity has important economic consequences.)

Emotions

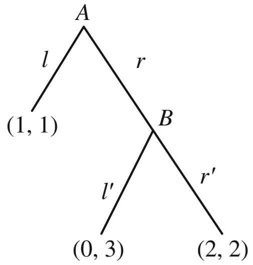
Reciprocity is but one form of motivation that can be modelled by means of psychological games. Many emotions are good candidates. In his article 'Emotions and Economic Theory', Elster (1998) argues that a variety of emotions have important economic consequences, and he laments how little attention economists have paid to this. He argues that a key characteristic of emotions is that 'they are triggered by beliefs' (Elster 1998, p. 49). He discusses anger, hatred, guilt, shame, pride, admiration, regret, rejoicing, disappointment, elation, fear, hope, joy, grief, envy, malice, indignation, jealousy, surprise, boredom, sexual desire, enjoyment, worry, and frustration. He asks (Elster 1998, p. 48): '[H]ow can emotions help us explain behavior for which good explanations seem to be lacking?' Psychological games may be useful for providing answers.

But little work has been done. One exception is Caplin and Leahy's (2004) health care model in which a physician is concerned with a patient's belief- dependent anxiety (compare also Caplin and Leahy 2001). Another exception is the emotion of guilt for which a string of results, both theoretical and experimental, have been established for the specific context of trust games (see Huang and Wu (1994), Dufwenberg (1995, 2002), Dufwenberg and Gneezy (2000), Bacharach, Guerra and Zizzo (Bacharach et al. 2007), and Charness and Dufwenberg (2006). I shall elaborate in some detail on these latter findings (borrowing eclectically from the cited works), since they may be suggestive of the importance of psychological games more generally in a variety of ways.

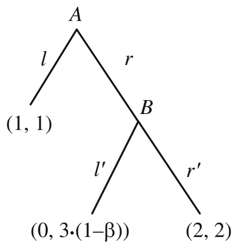
Consider the game in Fig. 1, where payoffs reflect money income (first for player *A*, then for player *B*) but not the players' preferences which may depend also on guilt as will be indicated.

Assume that the more strongly player *B* believes that player *A* believes that *B* will make choice r', the more guilt *B* would feel making choice l' and the more likely *B* is to make choice r'. Specifically, the players' utilities at the various end nodes in the game form of Fig. 1 coincide with the monetary payoffs, *except* following the choice sequence (r, l') where *B*'s utility is $3 \times (1 - \beta)$ rather than 3, and where β is a measure of B's belief (with range from 0 to 1) about A's belief that *B* will choose r'. (More specifically, *B* has a probability measure describing her beliefs about which probability *A* assigns to the choice r' conditional on *A* choosing r; β is the mean of that measure.) Say that *B* is *guilt averse.* This is all modelled in the psychological game in Fig. 2:

Psychological Games, Fig. 1



Psychological Games, Fig. 2



I wish to make several points. First, the guilt aversion modelled in Fig. 2 is similar to that involved in the above example featuring Karen and Jim. In fact, the idea that people feel guilty in proportion to the degree to which they do not live up to another's expectations can be extended to any game. (See Battigalli and Dufwenberg 2007, for a recent attempt at doing this.)

Second, one can test for guilt aversion experimentally, but this requires one to measure B's belief β. This can be done by inviting subjects to make guesses about one another's choices and guesses, rewarding accuracy in the guesswork. Such experimental tests have indicated that the prediction of guilt aversion is empirically supported in trust games. (The involved form of belief elicitation could conceivably be usefully complemented by two other forms of measurement: emotional selfreports and neurological methods such as functional magnetic resonance imaging.)

Third, guilt aversion may provide the seeds of a theory why communication can help foster trust and cooperation. To illustrate with reference to Fig. 2, suppose that, before play, *A* and *B* meet and talk. Player *B* looks player *A* in the eye and *promises* to choose *r*′. If *A* believes this, and if *B* believes that *A* believes this, then guilt aversion would make *B* live up to her promise. A promise by *B* can thus feed a self-fulfilling circle of beliefs about beliefs that *r*′ will be chosen. In combination with guilt aversion, words may be tools that create commitment power, which may in turn foster trust and cooperation.

Fourth, even without communication between *A* and *B*, one may argue that if *B* is guilt averse (as described above) then trust and cooperation will ensue. If *A* is rational and maximizes his expected monetary income (recall that we have assumed that *A* is selfish in this way) then by choosing *r* he *signals* a certain strength of belief in *B* choosing *r*′; if *A* did not assign a probability of at least 1/2 to *B* choosing *r*′ then he would rather chose *l*. If *B* figures this out, it puts a lower bound of 1/2 on β. So *B* is forced to hold a belief such that she would feel so guilty if she choose *l*′ that she prefers *r*′; in numbers, with $\beta \geq 1/2$ we get $3 \times (1 - \beta) < 2$. If *A* figures this out, he should of course choose *r*. The illustrated phenomenon has been labelled *psychological forward induction.*

To sum up: the idea of guilt aversion is intuitively plausible, experimentally testable, empirically supported, relevant for explaining why communication matters to economic behaviour, and suggestive of intriguing signalling issues that may shape emotions and behaviour. These insights concern a very special emotion and a very special psychological game, but seem profound given that limited scope. One may reasonably suspect that exciting conclusions are in store also for other emotions and other strategic settings, and that these conclusions may in part concern communication or belief signalling.

Social Rewards

The discussion so far may have been misleading with its rather heavy emphasis on reciprocity and emotions. Psychological game theory may be relevant also for describing certain social rewards (norms, respect and status), where decision makers somehow care about the opinions or views of others. Bernheim (1994) and Dufwenberg and Lundholm (2001) present models that bear this out. These authors do not make explicit mention of psychological games, but if one takes a close look at the mathematical details one can discover connections.

Developing the Theory

One might hope that Geanakoplos, Pearce and Stacchetti's framework is appropriate for tackling all the interesting problems to which psychological games may be relevant. However, this is not the case. A careful scrutiny reveals that their approach is too restrictive to handle many plausible forms of belief-dependent motivation (as they acknowledge themselves; see Geanakoplos et al. 1989, pp. 70, 78). There are several reasons, including the following:

(1) Geanakoplos, Pearce and Stacchetti only allow initial beliefs to enter the domain of a player's utility, while many forms of belief-dependent motivation require *updated* beliefs to play a role.
(2) Geanakoplos, Pearce and Stacchetti only allow a player's own beliefs to enter the domain of his utility, while there are conceptual and technical reasons to let *others'* beliefs matter.
(3) Geanakoplos, Pearce and Stacchetti follow the traditional extensive-games approach of letting strategies influence utilities only in so far as they influence which end node is reached, but many forms of belief-dependent motivation become compelling in conjunction with preferences that depend on strategies in ways not captured by end nodes.
(4) Geanakoplos, Pearce and Stacchetti restrict attention to equilibrium analysis, but in many strategic situations there is little compelling reason to expect players to coordinate on an equilibrium, and one may wish to explore alternative assumptions.

(1) is manifest, for example, in the above psychological forward induction argument which hinges crucially on B's motivation depending on an updated belief. (2) is relevant, for example, for modelling social rewards (compare the above comments on Bernheim's and Dufwenberg and Lundholm's models). As regards (3), one can show that the issue comes up if one wants to model, for example, regret, disappointment or guilt. (4) echoes considerations relevant also for traditional games; equilibrium play is not a self-evident proposition in many contexts, for example if one assumes (only) that there is common belief in rationality or in learning scenarios.

The list (1)–(4) is adapted from Battigalli and Dufwenberg (2005), who elaborate in more detail on each issue and take first steps towards developing psychological game theory in the indicated directions. Their approach draws crucially on Battigalli and Siniscalchi's (1999) work on how to represent hierarchies of conditional beliefs.

Decision-Theoretic Foundations

The decision-theoretic foundations of psychological game theory are not well understood. Classical decision theory (say, von Neumann and Morgenstern) does not apply straightforwardly. Too see this, take the emotion of disappointment as an example. It is plausible that disappointment is a belief-dependent emotion. To exemplify, I have I just failed to win a million dollars and I am not at all disappointed, which, however, I clearly would be if I were playing poker and knew I would win a million dollars unless my opponent got extremely lucky drawing to an inside straight, and then he hit his card. Another example could be based on the lotteries used in the so-called Allais paradox. In both cases the level of disappointment, which if anticipated might affect choice behaviour, may plausibly depend on the strength of the prior belief that a decision maker will win a lot of money. It follows that, unless consequences are described so as to include a specification of disappointment, the so-called 'independence axiom' will not make much sense for decision makers who are prone to disappointment.

Decision theorists have given related matters some attention, but not a lot. Machina (1981, pp. 172–3; 1989, p. 1662) presents examples in spirit related to the one million dollar example above. Bell (1985), Loomes and Sugden (1986), Karni (1992), and Karni and Schlee (1995) go on to develop models in which utility may depend directly on beliefs; the latter two references take axiomatic approaches. Robin Pope has written extensively, over many years, about how conventional decision theory excludes various forms of belief-dependent motivation; Pope (2004) expounds her programme and gives further references. Caplin and Leahy (2001) develop a model of 'psychological expected utility' that admits belief-dependent motivation. However, these contributions mainly develop perspectives for settings with single decision-makers, and more will be needed to address games more generally.

Conclusion

Research on psychological games is still in its infancy. This is true for all facets of investigation: the development of basic classes of games and solution concepts, the investigation of decision-theoretic underpinning, tests of empirical (most likely experimental) validity, and finally applied theoretical work which uses psychological game theory to analyse various economic models. In each of these domains some work has been done which is indicative of the viability of the line of research, and there is good reason to be thrilled about the prospects for future research.

See Also

- ▶ Behavioural Economics and Game Theory
- ▶ Behavioural Game Theory
- ▶ Expectations
- ▶ Expected Utility Hypothesis
- ▶ Game Theory
- ▶ Neuroeconomics
- ▶ Reciprocity and Collective Action
- ▶ Social Norms

Bibliography

Bacharach, M., G. Guerra, and D. Zizzo. 2007. The self-fulfilling property of trust: An experimental study. *Theory and Decision* 63: 349–388.

Battigalli, P., and M. Dufwenberg. 2005. Dynamic psychological games. Working Paper No. 287, IGIER, Bocconi University.

Battigalli, P., and M. Dufwenberg. 2007. Guilt in games. *American Economics Review* 97: 170–176.

Battigalli, P., and M. Siniscalchi. 1999. Hierarchies of conditional beliefs and interactive epistemology in dynamic games. *Journal of Economic Theory* 88: 188–230.

Bell, D. 1985. Disappointment in decision making under uncertainty. *Operations Research* 33: 1–27.

Bernheim, B.D. 1994. A theory of conformity. *Journal of Political Economy* 102: 841–877.

Caplin, A., and J. Leahy. 2001. Psychological expected utility and anticipatory feelings. *Quarterly Journal of Economics* 116: 55–79.

Caplin, A., and J. Leahy. 2004. The supply of information by a concerned expert. *Economic Journal* 114: 487–505.

Charness, G., and M. Dufwenberg. 2006. Promises and partnership. *Econometrica* 74: 1579–1601.

Dufwenberg, M. 1995. Time consistent wedlock with endogenous trust. Doctoral dissertation, *Economic studies 22*. Uppsala University.

Dufwenberg, M. 2002. Marital investment, time consistency, and emotions. *Journal of Economic Behavior and Organization* 48: 57–69.

Dufwenberg, M., and U. Gneezy. 2000. Measuring beliefs in an experimental lost wallet game. *Games and Economic Behavior* 30: 163–182.

Dufwenberg, M., and G. Kirchsteiger. 2004. A theory of sequential reciprocity. *Games and Economic Behavior 47*: *268–298*.

Dufwenberg, M., and M. Lundholm. 2001. Social norms and moral hazard. *Economic Journal* 111: 506–525.

Elster, J. 1998. Emotions and economic theory. *Journal of Economic Literature* 36: 47–74.

Fehr, E., and S. Gächter. 2000. Fairness and retaliation: The economics of reciprocity. *Journal of Economic Perspectives* 14: 159–181.

Geanakoplos, J., D. Pearce, and E. Stacchetti. 1989. Psychological games and sequential rationality. *Games and Economic Behavior* 1: 60–79.

Gilboa, Y., and D. Schmeidler. 1988. Information dependent games: Can common sense be common knowledge? *Economics Letters* 27: 215–221.

Huang, P., and H.-M. Wu. 1994. More order without more law: A theory of social norms and organizational cultures. *Journal of Law, Economics & Organization* 10: 390–406.

Karni, E. 1992. Utility theory with probability dependent outcome valuation. *Journal of Economic Theory* 57: 111–124.

Karni, E., and E. Schlee. 1995. Utility theory with probability dependent outcome valuation: Extensions and applications. *Journal of Risk and Uncertainty* 10: 127–142.

Loomes, G., and R. Sugden. 1986. Disappointment and dynamic consistency in choice under uncertainty. *Review of Economic Studies* 53: 271–282.

Machina, M. 1981. 'Rational' decision making versus 'rational' decision modeling. *Journal of Mathematical Psychology* 24: 163–175.

Machina, M. 1989. Dynamic consistency and non-expected utility models of choice under uncertainty. *Journal of Economic Literature* 27: 1622–1668.

Pope, R. 2004. Biases from omitted risk effects in standard gamble utilities. *Journal of Health Economics* 23: 1029–1050.

Rabin, M. 1993. Incorporating fairness into game theory and economics. *American Economic Review* 83: 1281–1302.

Rabin, M. 1998. Psychology and economics. *Journal of Economic Literature* 36: 11–46.
Sobel, J. 2005. Interdependent preferences and reciprocity. *Journal of Economic Literature* 43: 392–436.

Psychology and Economics

Charles R. Plott

Several developments have joined to stimulate economists to think about issues that have been on the forefront of psychological research. First, the information revolution in economics has focused economists on the subtle nature of individual information processing. Secondly, developments in game theory have so successfully identified new solution concepts that for almost any pattern of market behaviour there exists a reasonable theory consistent with that pattern. Introspection, a few principles of decision making, internal consistency, and a few stylized facts do not constrain possibilities enough to be sufficient guides to theory. Theorists are being forced to seek more systematic sources of data and additional principles to reduce the number of competing theories. Third, the rapid development of experimental methods applicable to economics has brought the testing of psychologically based economic theories within the realm of reality. Economists can accurately measure behaviour in economically relevant settings. As behavioural patterns become established that are difficult to reconcile with economic models alone, the profession has begun to look to psychology for answers. The data thus force the attention of economists to a broader class of models.

The interest of economists has also been stimulated directly by the work of psychologists whose own curiosity has brought their research closer to subjects that have been in the traditional domain of economics. Psychologists are deeply involved with risk/benefit analysis and normative decision analysis. Consequently, contact with traditional welfare economics and decision theory cannot be avoided. Psychologists have been actively studying commons dilemma problems. The literature has led them to an awareness of public goods, externalities, and notions of incentive compatibility. The trend is reversed in the study of non-human choices where economists interested in testing demand theory as applied to non-humans are being exposed to concepts of conditioning, shaping, reinforcement theory, etc., characteristic of the tools traditionally found useful to those who study animal behaviour in the laboratory.

A few exchanges between psychologists and economists have occurred recently. The focus of this section is on those exchanges. These are recent cases in which economists have become directly involved and have explored hypotheses that are of psychological origin. There is good reason to limit the material in this way. Almost any behaviour can have economic implications, and psychologists have not been particularly shy about making claims for the relevance of their research for economics. The whole field of psychology could be implicated. Consequently the focus will be on the work of economists that has been influenced by psychology as opposed to the work of psychologists that might have some economic import.

The Optimization Hypothesis

Economists tend to focus on choice and have had little to say about the process of choosing. The focus is natural. For the most part economics is about group phenomena and the behaviour of systems such as markets, whole economies, and political or other social units. The actions taken by each individual are the key contributions to the system behaviour and thus the individual actions are the important data. From this traditional perspective the particular cognitions that might have led to an act are of secondary importance to the action itself. The maintained hypothesis of goal seeking and purposeful actions as summarized by the optimization hypothesis provides a coherence and internal consistency to individual actions, but there has been no pressure to inquire deeper into the substance of the choosing process. The existence of attitudes is clearly acknowledged in the theory by the concepts of subjective probabilities and preferences, but the principles governing these attitudes have been maintained only at the most general level.

Choice is connected to the attitudinal parameters by a principle of optimization. The theory of rational choice claims only that observed choices will have that internal consistency necessary to be rationalized by a theory of optimization. Preferences will be revealed by choice.

There are many different types of rational choice. The choice can be influenced by random elements (McFadden 1974). The subjective probabilities and preferences can be related in ways not anticipated by the excessively strong expected utility hypothesis (Machina 1982; Chew 1983; Loomes and Sugden 1986). The preferences may not be transitive and instead only be devoid of cycles (Richter 1971). Only very recently has the literature seen generalizations of the concept of an optimum choice that does not require elements of maximization based on binary comparisons (Aizerman 1985). These principles are all of a most general nature and do not really address the issues of decision process because economic theory has evolved narrowly as demanded by the study of system behaviour.

By contrast, recent psychological research has been focused on the process of decisions as opposed to the choice or output of the process. The analysis reflects data generated by experimental methodology and is not constrained by systemic considerations. Perception, confusion, and the dynamics of preference and attitude formation are of central importance to psychologists.

Since the focus of psychological research is on decision processes, a natural tendency when confronted by the optimization hypothesis is to ask if people actually think in optimization terms. Do people consciously attempt to maximize something? The mathematics of optimization gives a rather clear outline of what a conscious process might be. Each option of a feasible set is compared in a series of binary comparisons of preference. The best is thereby identified and chosen. Do people do this? Do they *want* to do it?

Psychologists find little support for an optimization process as derived from the mathematical definitions. The lack of support for such a process is especially the case when uncertainty and information processing is involved. The natural response of the economics community when confronted by data that suggest that individuals do not consciously solve mathematical optimization problems has been to ask first, ‘so what?’. If the choices made by individuals naturally exhibit the internal consistency property of revealed preference then the cognitive processes are of minor consequence because the individuals would choose ‘as if’ optimizing and the substance of the theory of markets would be left intact. A second natural response of economists is to ask the psychologists if they have a theory that will ‘do better’ in explaining system behaviour. In the absence of an alternative theory, criticism of existing theory is not very compelling.

Psychologists have convincingly demonstrated that the ‘as if optimizing’ principle is not generally reliable. The easiest demonstration of the lack of generality of the principle involves the preference reversal phenomenon. The subject is given a choice between two lotteries. Lottery A is a .99 probability of winning \$4 and a .01 probability of winning nothing. Lottery B is a .33 probability of winning \$16 and a .67 probability of winning nothing. When asked to choose between the two, lottery A is the typical choice. When asked which lottery is valued the most, e.g., which of the lotteries has the highest reservation price or which has the highest selling price, then B is indicated. In other words the observed preference switches from A to B depending upon the way preference is measured. The phenomenon characterizes the behaviour of many people. It persists with monetary incentives and control for a variety of economic considerations (Grether and Plott 1979), training (Reilly 1982), and has been observed by a number of different researchers employing different techniques (Slovic and Lichtenstein 1983). The existence of the phenomenon is well documented.

The phenomenon is of interest to economists because it can be interpreted as an immediate intransitivity. Let $W =$ existing wealth, $\phi =$ the empty set, $\succsim$ indicate a preference relation, and $h(x)$ = the reservation price of lottery x. The pair (W, x) describes the possible states of an individual. The chain $(W + h(A), \phi) \sim (W,A) \succ (W,B) \sim (W + h(B), \phi)$ is observed from choice. Which do you prefer, A or B? Typically the subject chooses A as indicated. The minimum selling price when measured yields $h(A)$. Transitivity and the positive utility of money imply $h(A) > h(B)$, which contradicts the typical subject's expression of value; that B is valued more (has a higher reservation price) than A. The phenomenon is not only inconsistent with the expected utility hypothesis, it is inconsistent with all economic theories of preference. It demonstrates that choice is not universally consistent with an 'as if' optimization principle. Furthermore, since a single individual choosing among lotteries is a (degenerate) class of markets, the phenomenon bears directly on market choices.

Is there an alternative theory? The jury is still out on alternative explanations of preference reversals (Goldstein and Einhorn 1986; Loomes and Sugden 1982). However, the general approach to theories of decision processes taken by psychologists has been to retain the basic structure of optimization theory. People have transitive preferences but they might not be completely aware of them. Rather than ask what does a person want, the tendency of psychologists is to ask how a person *expects* to feel as a result of taking an act (March 1978). Perception, limited capacities for calculation, and perhaps the dependence of preferences on previous choices are integral aspects of the modified theory.

Choice is preceded by an editing phase of activities. Alternatives are first examined for attributes, which are coded as a gain or loss relative to some reference point. The reference point could be a status quo or it could be a level of aspiration. This editing phase is sometimes referenced as 'framing' (Tversky and Kahneman 1981).

The evaluation phase follows editing. Almost all psychological theories of the process of choice postulate that the individual compares options attribute by attribute. This is in contrast to a holistic assessment in which the individual evaluates all attributes of a single option and, having assessed the option, the process moves to another option which is wholly assessed. Studies that involve the tracing of eye movements during the process of choice support the psychologists' presumptions (Russo and Dosher 1983). Eye movement tends to fix on an attribute and move across options, remaining fixed on that attribute. After surveying objects, a new attribute is chosen. This process should be contrasted to one in which all attributes of an object are examined as if to gain a total assessment or utility of the object before moving to the next object.

Slovic and Lichtenstein (1983) think of evaluation in terms of anchoring and adjustment. This process begins with the identification and evaluation of a single attribute. The value of that attribute is attached to the option and then the valuation is adjusted upward or downward to compensate for other attributes. Inconsistencies such as those observed in preference reversals are the result of incomplete adjustments according to this model. Risk averse people tend to choose option A when the choices are offered. Then, when such people are asked to place a value on the options, they focus first on the monetary amount, attach that initial value to the object, and then adjust the amount downward to compensate for the probability. Incomplete adjustment for the probability and other attributes yields a higher value for B than for A, which began with a lower initial valuation because of the lower dollar figure from which base adjustments are made.

The result is the observed inconsistency with choice. Other evaluation models exist. Lexicographic rules involve an ordering of attributes. The chosen alternative has the highest level of the best attribute. Prospect theory (Tversky and Kahneman 1981) involves valuing attributes but the theory incorporates an asymmetry between losses and gains. When the options are lotteries, prospect theory holds that people are risk avoiders in gains and risk seekers in losses. Of course, since lexicographic rules and prospect theory are optimization theories, neither can account for the preference reversal phenomena.

Another class of models are suggestive of what Simon (1955, 1979) has called satisficing. The choosing agent is viewed as being involved in a costly search problem as objects are examined and attributes are compared. The theoretical problem becomes one of isolating the rules governing search and decision. Two important classes of such rules are conjunctive and disjunctive (Einhorn 1970). Conjunctive and disjunctive rules require cutoff levels be set on each attribute. A conjunctive rule requires that any option with an attribute below cutoff is rejected. Disjunction requires that options with attributes above cutoff be accepted.

Economists have pursued some aspects of this theory (Grether and Wilde 1984). Optimal risk-neutral conjunction decision rules were calculated by the researchers for an experimental economics setting involving sequential (search) decisions.In addition, a set of nonoptimal rules were calculated. The non-optimal rules were approximations of the optimal rules in which the cross derivative terms in the first-order conditions of an optimal sequential decision were ignored. Subjects were restricted to the use of conjunctive rules so the test was whether or not consumers used optimal rules or the postulated non-optimal rules given that a particular type of decision rule had to be used. Subjects tended to use the nonoptimal rules in which 'second-order' tradeoffs are ignored.

The Grether and Wilde study has two major dimensions of interest. First, their theoretical methods incorporate a way to identify 'rules of thumb' that have a systematic and theoretical departure from optimal. They have a way of using optimal rules to generate simplifications of the rule that can be interpreted as satisficing. Secondly, the experimental methods provide a way of testing such modified theories and getting good measurements on the nature of their inadequacies.

Bayes' Law

Economic models tend to treat individuals as if they were intuitive statisticians. Of course Bayes law is a central tenet. It is used in the theory without modification any time an economic agent is exposed to information.

By contrast, the psychological approach to subjective probability is similar to the psychological approach to decisions in general. People are viewed as using rules of thumb which may or may not reflect an appropriate statistical principle.

One of these rules called the representativeness heuristic has caught the attention of economists. According to this rule individuals will view samples as having come from the population that is most representative of the sample. The rule suggests a tendency to ignore prior probabilities or, in a sense, put too much weight on the sample while generating a posterior. The phenomenon has been studied extensively in the psychology literature. The following is an example of what is observed. Suppose urn A contains black balls and white balls with $Pr(black) = .75$ and $Pr(white) = .25$. Urn B has the opposite probabilities. A sample of four balls with replacement will be drawn from one of these urns. The subject must guess the urn that was used and is rewarded cash for correct choices. The urn to be used in the draw is decided from a draw from a third urn with $Pr(A) = .01$ and $Pr(B) = .99$.

Now suppose a sample of four balls is drawn. Three of the four are black. Which urn was used?

Bayes law gives $Pr\ (A) \approx 0.02$ Nevertheless, the representativeness hypothesis predicts that A will be chosen. The reason is the similarity between the sample and the distribution of balls in A. The prior probabilities will be underweighted according to this theory.

An experiment conducted by Grether (1980) was of the form used in the example. Rewards provide a financial incentive to choose the urn that the subject thinks is the most likely. That is, subjects are paid for guessing the correct urn. For a given sample, x, urn A is chosen when $Pr\ (A\,|\,x)/Pr\ (B\,|\,x) > 1$ and B is chosen otherwise.

The use of Bayes law gives the model

$$\frac{Pr(A|x)}{Pr(B|x)} = \left[\frac{Pr(x|A)}{Pr(x|B)}\right]^{\beta_1}\left[\frac{Pr(A)}{Pr(B)}\right]^{\beta_2}$$

where $\beta_1 = \beta_2 = 1$ and $Pr(A)$ and $Pr(B)$ are prior probabilities. Letting Y^*_{it} be the subjective log odds in favor of A for person i at time t so $e^{Y^*_{it}}$ is the posterior odds in favor of A. By taking the logarithm of

$$e^{Y^*_{it}} = e^{\alpha}\left[\frac{Pr(x|A)}{Pr(x|B)}\right]^{\beta_1}\left[\frac{Pr(A)}{Pr(B)}\right]^{\beta_2} et^{u_{it}},$$

where u_i is a random variable, a model suitable for estimation by logit is obtained. Of course Y^*_{it} is not observed but the observed choice provides a variable $Y_{it} =$ when $Y^*_{it} \geq 0$ and $Y_{it} = 0$ otherwise. Bayes law yields the hypothesis $\alpha = 0$, $\beta_1 = \beta_2 > 0$ and the representatives hypothesis, $\beta_1 > \beta_2 \geq 0$.

The qualitative predictions of the representativeness hypothesis have been supported by the experimental data. However, the coefficient β_2 tends to be strictly greater than zero so the prior probabilities are not ignored. Generally speaking the Bayes law model has very good predictive ability. Bayes law is not ‘bad’ as a model but its predictive power is increased by allowing for the possibility of a representativeness heuristic.

Conclusion

When examining psychologically motivated theories, economists have differed from psychologists in three methodological ways. First, economists have tended to use as the dependent variables an act or an observed choice. By contrast, psychologists have tended to elicit attitudes, such as a degree of preference, belief, or numerical probability. Secondly, economists have been more sensitive to the possibility that motivated choice might differ from unmotivated choice. The third difference is the context. Economists have tended to avoid the rich descriptive hypothetical contexts and scenarios used by psychologists. This no doubt reflects a fear that the verbal descriptions will influence behaviour.

For the most part, when testing directly, economists have observed for themselves what psychologists claimed they would observe. Incentives do matter but so far incentives have not been observed overriding the tendencies that psychologists have observed when incentives were absent. It is too early to speculate about what future research will uncover regarding this issue.

Whether or not the psychologically based theories will account for significant market behaviour remains to be established. The potential for such theories has been recognized (Thaler 1980) but so far the psychological theories, when applied to experimental markets, have not been sufficiently supported to suggest the need for a substantial overhaul in economic theory (Plott 1986). Nevertheless, the field has only just begun and, given the persistence and magnitude of effects uncovered at the individual level of analysis, it is reasonable to expect that the effects will soon be detected in experimental markets.

See Also

- ▶ Experimental Methods in Economics
- ▶ Political Economy and Psychology
- ▶ Preference Reversals

Bibliography

Aizerman, M.A. 1985. New problems in the general choice theory: Review of a research trend. *Social Choice and Welfare* 2: 235–282.

Chew, S.H. 1983. A generalization of the quasilinear mean with applications to the measurement of income inequality and decision theory resolving the Allais paradox. *Econometrica* 51: 1065–1092.

Einhorn, H. 1970. The use of nonlinear, noncompensatory models in decision making. *Psychological Bulletin* 73: 221–230.

Goldstein, W.M., and H. Einhorn. 1986. Expression theory and the preference reversal phenomena. *Psychological Review* 89: 627.

Grether, D.M. 1980. Bayes’ rule as a descriptive model: The representativeness heuristic. *Quarterly Journal of Economics* 95: 537–557.

Grether, D.M., and C.R. Plott. 1979. Economic theory of choice and the preference reversal phenomenon. *American Economci Review* 69: 623–638.

Grether, D.M., and L.L. Wilde. 1984. An analysis of conjunctive choice: Theory and experiments. *Journal of Consumer Research* 10: 373–385.

Loomes, G., and R. Sugden. 1982. Regret theory: An alternative theory of rational choice under uncertainty. *Economic Journal* 92: 805–824.

Loomes, G., and R. Sugden. 1986. Disappointment and dynamic consistency in choice under uncertainty. *Review of Economic Studies* 53: 271–282.

Machina, M.J. 1982. Expected utility analysis without the independence axiom. *Econometrica* 50: 277–323.

March, J.G. 1978. Bounded rationality, ambiguity and the engineering of choice. *Bell Journal of Economics* 9: 587–610.
McFadden, D. 1974. Conditional logit analysis of qualitative choice behavior. In *Frontiers in econometrics*, ed. P. Zarembka, 105–142. New York: Academic Press.
Plott, C.R. 1986. Rational choice in experimental markets. *Journal of Business* 59(2): S301–S334, Pt. 2.
Reilly, R.J. 1982. Preference reversal: Further evidence and some suggested modifications in experimental design. *American Economic Review* 72: 576–584.
Richter, M.K. 1971. Rational choice. In *Preference, utility and demand*, ed. J.S. Chipman et al. San Francisco: Harcourt, Brace, Jovanovich.
Russo, J.E., and B.N. Dosher. 1983. Strategies for multiattribute binary choice. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 9: 676–696.
Simon, H.A. 1955. A behavioral model of rational choice. *Quarterly Journal of Economics* 69: 99–118.
Simon, H.A. 1979. Rational decision making in business organizations. *American Economic Review* 69: 493–513.
Slovic, P., and S. Lichtenstein. 1983. Preference reversals: A broader perspective. *American Economic Review* 73: 596–605.
Thaler, R. 1980. Toward a positive theory of consumer choice. *Journal of Economic Behavior and Organization* 1: 39–60.
Tversky, A., and D. Kahneman. 1981. The framing of decisions and the psychology of choice. *Science* 211: 453–458.

Psychology of Social Networks

Philippa Pattison, Garry Robins and
Yoshi Kashima

Abstract

We review the psychological processes underpinning network formation and network-based processes, focusing first on the nature of relationships and their formation, and then on the consequences of networks for individual outcomes and behavior. We argue that it is important to develop methodological approaches that allow us to regard these processes as hypotheses to be tested rather than as unquestioned assumptions. We suggest that different types of networks and processes are likely to lead to different conclusions about these hypotheses, and that the development of models for networks and network processes should therefore be grounded in careful empirical analysis.

Keywords

Social networks; Psychology of; Network formation; Network models; Network ties; Relational models; Relational schemas; Social roles; Tie interdependencies; Exponential random graph models

JEL Classifications

D85

The significance of social networks for an understanding of the structure and dynamics of our contemporary social world is now widely acknowledged. Different types of networks serve as channels through which, for example, knowledge is diffused, opportunities are recognized, cooperation is garnered and actions are coordinated. Networks have been invoked in many disciplines in order to explain the nature and consequences of these channelling effects, spawning interest in the capacity to model network structure. This capacity is important because of the potentially powerful interplay between network structures and the dynamics of the social transactions that they support.

Two broad strategies for network model building have been identified (Jackson 2005): a statistical approach, in which networks are seen as the outcome of locally interactive and self-organizing tie-formation processes; and a deterministic, game-theoretic approach, in which networks are seen as the outcome of self-interested behaviour of utility-maximizing actors. In association with the statistical approach, there has been a dramatic increase in our capacity to build theoretically defensible statistical models for social networks whose parameters can be estimated from empirical observations and that have the capacity to reproduce many important global characteristics of observed social networks (for example, Snijders et al. 2006a, b). The game-theoretic approach, on the other hand, is responsible for an impressive accumulation of theoretical results linking strategic activities of pairs of actors to the emergence of 'efficient' network structures.

These two modelling approaches differ in two important respects. The first is in the use of deterministic or stochastic models (but see Snijders 2001). The second distinction is a deeper one, contrasting a conceptualization of actors as self-interested rational decision makers, on the one hand, with more socialized and enculturated actors, on the other. The latter may sometimes be driven by personal utility, but may also engage in non- or extra-rational behaviours that are enabled and constrained by local social network configurations. Indeed, whereas the game-theoretic approach largely assumes that the actor's decision to form or discontinue a relationship is mediated by the actor's conscious computation of utility, the statistical approach treats this as a hypothesis, and empirically examines *whether* an actor's local network configuration appears to constrain or enable tie formation or maintenance. Thus, this second distinction encapsulates a fundamental empirical question: can network structures and processes be explained in terms of the rational activity of self-interested individuals, or are extra-rational affective, social and/or cultural processes systematically at work in the formation and impact of networks?

With this question in mind, we outline what is known about the social and psychological processes that underpin the formation of networks and the dynamics of network-based processes. We first discuss the nature of relationships and their development, and then examine their consequences for individual outcomes and behaviour. We finish by drawing some implications for model-building.

What is a Network Tie and Why Do Relationships Form?

A network relationship, or tie, is assumed to have some continuity in time, with a relevant past and a somewhat predictable future. This temporal continuity is facilitated by cognitive representation and the use of culturally laden relational descriptors, such as friend and friendship, or partner and partnership, features that also facilitate communication about the nature of ties. Relationships, in other words, entail complex socio-cultural schema that not only frame interpretations of past interactions but also shape expectations concerning future ones by the actors concerned as well as by third-party onlookers.

Although tie formation has sometimes been explained in terms of the utility that ties bring for tie partners (for example, Cook and Whitmeyer 1992), recent psychological research suggests that significant extra-rational psychological processes may also underlie tie formation. Anderson and Chen (2002) have invoked the concept of *relational self* to explain the pervasive impact of relationships with significant others on the way in which individuals interpret and respond to interpersonal encounters, and therefore form *future* interpersonal ties. This concept is founded on a demonstrable 'transference' effect, in which past experiences with significant others can be shown to influence new relationships, often outside of conscious awareness. Holmes (2000) has argued that the concept of relationship is best seen as grounded in the interaction between interdependent actors, and hence as an emergent property of dynamic interaction and influence processes. Generic cognitive representations about relationships called *relational schemas* are postulated to represent actors' developing knowledge of self, partner and expected sequences of interaction (for example, Baldwin 1992).

More generally, Fiske (2004) has proposed that four elementary and universal cognitive schemas frame *all* interpersonal relationships. These four schemas are proposed to structure potential interactions between two actors in terms of: collective belonging or solidarity, as in family membership (the *communal sharing* schema); asymmetrical difference, as in hierarchies based on skill, knowledge or social class (*authority ranking*); an egalitarian relationship, based on turn-taking and exchange, as in many friendship ties (*equality matching*); and a rational analysis of costs and benefits, as in a payment-for-service regime (*market pricing*). Fiske claims that any actual social relationship is constituted by some mixture of these forms, and that socially transmitted interpretive guidelines link these universal forms to specific relationship characteristics in a particular culture.

While relational schemas have advanced our understanding of the nature and variety of dyadic relationships, there is a recognized need to understand the specific ways in which they depend on social situations and give rise to interdependencies among relationships *across* a network. As Haslam (2004, p. 297) has observed, Fiske's relational models theory posits 'a universal grammar of social relations . . . out of whose rules and representations the myriad local forms of social life can be generated'. Haslam argues that the categorical nature of the relational models may underpin the complexity of human social organization by facilitating the required coordination of interpersonal obligations, rights and responsibilities. Social roles are important intermediate-level constructs in this account, and are seen as 'distinctively implemented admixtures' of relational models, a view that is supported by evidence that relational models mediate the effects of social roles on social cognition. More generally, social roles have been invoked by many theorists to explain interdependencies among multiple relationships across multiple actors (White et al. 1976).

It follows from these claims that tie formation processes should depend on the type of tie under consideration. While some ties may be consciously negotiated at a dyadic level, independently of other ties, others may be subject to subtle influences arising from the embedding of the potential tie in a local social setting that comprises ties to and among third-party actors with their own mix of potentially competing and potentially cooperative goals. As a result, Fiske's communal sharing relations, for example, should exhibit much stronger interdependencies across pairs of relation partners than relations of the market pricing kind. Such effects have been empirically demonstrated. For example, Granovetter's influential hypothesis concerning the 'strength' of weak ties was based on the distinction between the highly clustered structure of 'strong' tie networks and the less structured, more open, spreading character of 'weak' tie networks, an hypothesis that has received empirical support (Granovetter 1982). There is also evidence that different types of network tie can be mutually interdependent and subject to generalized forms of exchange and interlock through ties to third parties (for example, Lomi and Pattison 2006). Indeed, changes in patterns of cross-network interdependence have been invoked in explanations of social and economic innovation (for example, Padgett and McLean 2006).

In addition to the interdependencies just described, there is also compelling evidence for the impact of individual actors' characteristics on the formation of network ties. Tie partners are more likely to share socio-demographic characteristics such as gender, age, ethnicity and religion (for example, McPherson et al. 2001). The formation of relationships is also clearly a function of social settings that affect the probability of any two actors having an opportunity to interact (Feld 1981). The psychological literature on relationship formation emphasizes the importance of more psychological similarities among potential tie partners, a premise that has been extended by Robins and Boldero (2003) to include a comparison of potential tie partners' aspirations and obligations.

Taken together, these structuring influences on tie formation can be seen as operating at multiple levels, with broad socio-demographic factors at work on a larger scale, and more micro-social and psychological factors at work at more local scales. Whereas the broader factors can often be regarded as exogenous influences on tie formation, the more micro-level factors are usually best seen as endogenous, with the tie formation processes for one pair of actors having consequences for tie formation among their network partners, and their partners, and so on. While some of these interdependencies may operate outside the awareness of individual actors, there are also circumstances in which actors seek out institutional settings and particular relationships precisely for the strategic 'networking' opportunities that they provide. Moreover, expected interdependencies can themselves be countered: Padgett and Ansell (1993) coined the term 'robust action' to describe behaviour that is open to multiple interpretations, and therefore has the capacity to elide third-party influences on tie formation.

How Do Relationships Affect Individual Outcomes?

An actor's location in a social network is an important aspect of social context and hence potentially plays an important role in determining many types of future behaviour apart from the development of social ties. There are a number of related mechanisms by which such effects might occur (for example, Pattison 1994). First, network ties serve as a conduit for information, and hence specific network locations can have a dramatic impact on the information or other resources that any one actor possesses (for example, Burt 2004). This information can itself be subject to subtle filtering effects, such as the suppression of information perceived to be inconsistent with shared understanding at a community level (Lyons and Kashima 2003). Second, network ties can influence actors' understanding of relationships among others as well as expectations about the future behaviour of others. For example, they are likely to be more certain (and more accurate) in judging the relationships involving their network partners and, to a lesser extent, their network partners' partners (Kumbasar et al. 1994). Finally, social influence effects may be brought to bear as actors weigh up – consciously or unconsciously – the views of others in forming or modifying their own beliefs (for example, Friedkin 1998; Robins et al. 2001).

Implications for Model-Building

Models for network structure and network evolution almost certainly need to accommodate many of the exogenous and endogenous influences on tie formation just described. An appropriate model class is the exponential family of random graph models that was first introduced by Frank and Strauss (1986), building on the general formulation of statistical models for interacting systems of variables by Besag (1974). The use of principled approaches to specifying potential tie interdependencies (Pattison and Robins 2002) has led to models that yield impressive fits to even large observed network structures (for example, Goodreau 2007). Interestingly, these models combine a Markov dependence assumption (that potential network ties with an actor in common are dependent, conditional on the state of all other potential ties in a network) with a 'longer range' form of assumed dependence (Snijders et al. 2006a, b) in which, in some circumstances, ties involving discrete pairs of actors are also conditionally dependent. The necessity of these assumptions in many empirical contexts (Robins et al. 2006) suggests that interdependencies among ties within local social contexts are indeed an important influence on observed network forms.

Models for individual states and choices (including beliefs and actions) may likewise need to accommodate subtle network-based interdependencies among the states and choices of their tie partners. Robins et al. (2001) have developed a general social influence modelling framework for this purpose, akin to the network modelling framework just mentioned.

It is important in the context of the question posed earlier – whether network structures and processes can be explained in terms of the rational activity of self-interested individuals, or whether there are extra-rational processes at work in the formation and impact of networks – to develop observational designs and analytic methods that allow us to regard tie formation and network processes as hypotheses to be tested rather than unquestioned assumptions. We might speculate that different types of networks and different types of social processes are likely to lead to different conclusions about these guiding hypotheses, and that, as a consequence, models should continue to be developed from multiple perspectives and to be grounded in careful empirical analysis. Finally, it is worth noting that the development of methods for estimating models from longitudinal observations may prove particularly helpful in sifting among alternative approaches to model building, not just for models of network evolution and social influence dynamics but also for new approaches to modelling the co-evolution of networks and behaviour (Snijders et al. 2006a, b).

See Also

- ▶ Emergence
- ▶ Learning and Information Aggregation in Networks
- ▶ Mathematics of Networks
- ▶ Network Formation
- ▶ Psychology of Social Networks
- ▶ Social Interactions (Empirics)
- ▶ Social Interactions (Theory)
- ▶ Social Networks in Labour Markets

Bibliography

Anderson, S., and S. Chen. 2002. The relational self: An interpersonal social-cognitive theory. *Psychological Review* 109: 619–645.

Baldwin, M. 1992. Relational schemas and the processing of social information. *Psychological Bulletin* 112: 461–484.

Besag, J. 1974. Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society, Series B* 36: 96–127.

Burt, R. 2004. Structural holes and good ideas. *American Journal of Sociology* 110: 349–399.

Cook, K., and J. Whitmeyer. 1992. Two approaches to social structure – exchange theory and network analysis. *Annual Review of Sociology* 18: 109–127.

Feld, S. 1981. The focused organization of social ties. *American Journal of Sociology* 86: 1015–1035.

Fiske, A. 2004. Relational models theory 2.0. In *Relational models theory: A contemporary overview*, ed. N. Haslam. Hillsdale: Lawrence Erlbaum.

Frank, O., and D. Strauss. 1986. Markov graphs. *Journal of the American Statistical Association* 81: 832–842.

Friedkin, N. 1998. *A structural theory of social influence*. New York: Cambridge University Press.

Goodreau, S. 2007. Advances in exponential random graph (p*) models to a large social network. *Social Networks* 29: 231–248.

Granovetter, M. 1982. The strength of weak ties: A network theory revisited. In *Social structure and network analysis*, ed. P. Marsden and N. Lin. Beverly Hills: Sage.

Haslam, N. 2004. Four grammars for primate social relations. In *Evolutionary social psychology*, ed. J. Simpson and D. Kenrick. Hillsdale: Lawrence Erlbaum.

Holmes, J. 2000. Social relationships: The nature and function of relational schemas. *European Journal of Social Psychology* 30: 447–496.

Jackson, M. 2005. 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. New York: Cambridge University Press.

Kumbasar, E., A. Romney, and W. Batchelder. 1994. Systematic biases in social perception. *American Journal of Sociology* 100: 477–505.

Lomi, A., and P. Pattison. 2006. Manufacturing relations: An empirical study of the organization of production across multiple networks. *Organization Science* 17: 313–332.

Lyons, A., and Y. Kashima. 2003. How are stereotypes maintained through communication? The influence of stereotype sharedness. *Journal of Personality and Social Psychology* 85: 989–1005.

McPherson, M., L. Smith-Lovin, and J. Cook. 2001. Birds of a feather: Homophily in social networks. *Annual Review of Sociology* 27: 415–444.

Padgett, J., and C. Ansell. 1993. Robust action and the rise of the Medici, 1400–1434. *American Journal of Sociology* 98: 1259–1319.

Padgett, J., and P. McLean. 2006. Organizational invention and elite transformation: The birth of partnership systems in renaissance Florence. *American Journal of Sociology* 111: 1463–1568.

Pattison, P. 1994. Social cognition in context: Some applications of social network analysis. In *Advances in social network analysis in the social and behavioral sciences*, ed. S. Wasserman and J. Galaskiewicz. Newbury Park: Sage.

Pattison, P., and G. Robins. 2002. Neighbourhood-based models for social networks. *Sociological Methodology* 32: 300–337.

Robins, G., and J. Boldero. 2003. Relational discrepancy theory. *Personality and Social Psychology Review* 7: 56–74.

Robins, G., P. Pattison, and P. Elliott. 2001. Network models for social influence processes. *Psychometrika* 66: 161–190.

Robins, G., T. Snijders, P. Wang, M. Handcock, and P. Pattison. 2006. Recent developments in exponential random graph (p*) models for social networks. *Social Networks* 29: 192–215.

Snijders, T. 2001. The statistical evaluation of social network dynamics. In *Sociological methodology 2001*, ed. M. Sobel and M. Becker. Boston/London: Basil Blackwell.

Snijders, T., C. Steglich, and M. Schweinberger. 2006a. Modeling the co-evolution of networks and behavior. In *Longitudinal models in the behavioral and related sciences*, ed. K. van Montfort, H. Oud, and A. Satorra. Hillsdale: Lawrence Erlbaum.

Snijders, T., P. Pattison, G. Robins, and M. Handcock. 2006b. New specifications for exponential random graph models. *Sociological Methodology* 36: 99–153.

White, H., S. Boorman, and R. Breiger. 1976. Social structure from multiple networks: I. Blockmodels of roles and positions. *American Journal of Sociology* 87: 517–547.

Public Capital

David A. Aschauer

Abstract

This article reviews recent research emphasizing the potential importance of public capital (or infrastructure) to aggregate economic performance, and provides a survey of empirical estimates of the productivity of public capital and of the impact of public capital investment on economic growth.

Keywords

Cobb–Douglas functions; Endogenous growth; Long-term economic growth; Neoclassical growth theory; Output elasticity of public capital; Productivity growth; Public capital; Public finance; Returns to public capital; Total factor productivity

JEL Classifications

H0; E21

Public capital (or often 'infrastructure') encompasses the publicly provided capital facilities which form the basis for private sector economic activity.

Empirically, public capital typically is defined as a net (of depreciation) stock of non-military structures and equipment and is often decomposed into *core* public capital (consisting of transportation facilities – such as streets and highways, mass transit, rail, and airports, water and sewer systems, and electrical and gas facilities), and *other* public capital (comprising educational structures, public hospitals, courthouses and the like).

The Productivity of Public Capital

Beginning at the end of the 1980s, a significant research effort has focused on estimating the contribution of public capital to macroeconomic performance. The research initiative seems to have been the result of the recognition of certain facts about public capital expenditures in the United States. First, infrastructure capital accumulation, when expressed as a fraction of output, began to decline toward the end of the 1960s and, as a result, was seen as a potential factor in explaining the productivity growth slowdown of the 1970s and 1980s. Second, during the same period, the United States devoted a smaller share of output to infrastructure than did other industrialized economies (such as those in the Group of Seven), which was taken as possible force in explaining the relatively low rate of productivity growth in the United States vis-à-vis other countries such as Japan and Germany.

The first stage of the research effort centred on estimating the contribution of infrastructure to private sector productivity, where infrastructure is taken as another factor of production, along with private capital and labour, in an aggregate production function of the form

$$Y = A \cdot F\left(l, K, K^G\right)$$

where Y denotes the aggregate level of economic output, A an index of total factor productivity, L, the labour force or employment, K private capital (usually restricted to business fixed capital), and K^G the stock of public capital. The basic goal of the research was to ascertain the value of the output elasticity of public capital

$$\varepsilon = \frac{K^G}{Y} \cdot \frac{\partial Y}{\partial K^G}$$

in order to determine the 'productivity of public capital'.

The early empirical results, typically employing level data in estimating a Cobb–Douglas production function, indicated (strikingly) high elasticity estimates, in the range of 0.25 to 0.50 for the United States and even higher for countries such as Canada and Sweden. These elasticity estimates, in turn, implied very high rates of return to public capital investments which some took as implausible. For example, Gramlich (1994) used Aschauer's (1989) elasticity estimate of 0.39 to generate an estimate of the marginal productivity of public capital in the range of 0.70 to 1.00, which, in his view, was implausible since it implied that investments in government capital generate enough extra output to pay for themselves in a year.

Later, a number of researchers estimated the production function using first-differenced data, arguing that the initial results were 'spurious' because (*a*) variables such as output and public capital were first-order integrated series and (*b*) the production function did not serve as a cointegrating relation between output and the various factors of production (including public capital). These studies (for example, Tatom 1991) often generated much lower, and less reliable, estimates of the output elasticity of public capital.

Recently, Kamps (2004) has developed new estimates of public capital stocks for 22 OECD countries over the period 1960–2001, and has estimated the output elasticity of public capital. The point estimates are positive in 20 of 22 cases and statistically significant in 12 of 22 cases. A panel regression employing first-differenced data leads to a reasonable elasticity estimate equal to 0.22 which leads the author to conclude that public capital is productive on average in the countries comprising the OECD.

Public Capital and Economic Growth

The finding that public capital is productive is not, in and of itself, evidence that increasing public capital investment will raise long-term economic growth. There are at least three considerations which must be addressed. First, there is the question of whether a permanent increase in public investment will induce a permanent or transitory increase in growth. The traditional neoclassical growth model predicts that an increase in national savings and investment rates will have only a transitory effect on growth; more recent endogenous growth models, on the other hand, would predict permanent effects. Second, given the level of national savings, the effect of public investment on economic growth depends not just on a positive output elasticity of public capital, but on the relative marginal productivities of private and public capital; an increase in public investment at the expense of public investment will raise or lower the economic growth rate depending on whether the marginal product of public capital exceeds, or is exceeded by, the marginal product of private capital. Third, the effect of public capital on economic growth will depend on the method of public finance – whether by current taxes, debt, or (potentially) money creation.

One approach which allows tentative answers to all three questions is that of Aschauer (2000), who extends the Barro (1990) model of productive government spending to explicitly include public investment. This model, which assumes (*a*) that public investment is debt-financed and (*b*) a production function which displays constant returns to scale across private and public capital stocks (per worker) generates endogenous growth in per worker output at the rate

$$\gamma_y = \frac{1}{\sigma} \cdot \left[(1-\tau) \cdot (1-\varepsilon) \cdot \left(\frac{K^G}{K} \right)^{\varepsilon} - \rho \right]$$

where y is the level of output per worker, $(1/\sigma)$ the intertemporal elasticity of substitution, τ the tax rate necessary to service the public debt associated with public capital, and ρ a rate of time preference. Evidently, increases in the tax rate lower economic growth, while increases in the ratio of public capital to private capital raise economic growth. It turns out that increases in public capital will raise or lower economic growth depending on whether the tax rate is lower or higher than the output elasticity of public capital – that is, there is a nonlinear relationship between public capital and growth and an 'optimal' level of public capital. Using US state level data, Aschauer finds robust evidence that the relationship between public capital and growth is, indeed, nonlinear and that public capital is underprovided – that is, the 'optimal' ratio of public capital to private capital is in the range of 0.60 while the actual average ratio equals 0.44. As a consequence, a ten per cent increase in the public capital ratio is estimated to raise economic growth by approximately one percentage point per year.

See Also

- ▶ Monetary and Fiscal Policy Overview
- ▶ Optimal Fiscal and Monetary Policy (Without Commitment)

Bibliography

Aschauer, D.A. 1989. Is public expenditure productive? *Journal of Monetary Economics* 23: 177–200.

Aschauer, D.A. 2000. Do states optimize? Public capital and economic growth. *Annals of Regional Science* 34: 343–364.

Barro, R.J. 1990. Government spending in a simple model of endogenous growth. *Journal of Political Economy* 98: 103–125.

Gramlich, E.M. 1994. Infrastructure investment: A review essay. *Journal of Economic Literature* 32: 1176–1196.

Kamps, C. 2004. New estimates of government net capital stocks for 22 OECD countries 1960–2001. Working paper 04/67, International Monetary Fund.

Tatom, J.A. 1991. Should government spending on capital goods be raised? *Federal Reserve Bank of St. Louis Review* 73(2): 3–15.

Public Choice

Gordon Tullock

Abstract

By assuming that voters, politicians and bureaucrats are mainly self-interested, public choice uses economic tools to deal with the traditional problems of political science. Its findings revolve around the effects of voter ignorance, agenda control and the incentives facing bureaucrats in sacrificing the public interest to special interests. The design of improved governmental methods based on the positive information about how governments actually function has been an important part of public choice. Constitutional reforms advocated variously by public choice thinkers include direct voting, proportional representation, bicameral legislatures, reinforced majorities, competition between government departments, and contracting out government activities.

Keywords

Agenda control; Arrow, K.; Bicameral legislatures; Black, D.; Buchanan, J.; Bureaucracy; Constitutions, economic approach to; Democracy; Direct voting; Downs, A.; Niskanen, W.; Proportional representation; Public choice; Public interest; Reinforced majorities; Single-peaked preferences; Social choice; Special interests; Tullock, G.; Voting

JEL Classifications

H0

In the 18th and 19th centuries a number of mathematicians (Condorcet, Borda, Laplace and Lewis Carroll) became interested in the mathematics of the voting process; their work was forgotten until Duncan Black rediscovered it (see, e.g., Black 1958). Black can be called the father of modern Public Choice, which is in essence the use of economic tools to deal with the traditional problems of political science. Historically, economics (political economy) dealt to a very large extent with the choice of government policies with respect to economic matters. Whether protective tariffs were or were not good things would be a characteristic topic of traditional economics and in examining the question, it was assumed, of course, that the government was attempting essentially to maximize some kind of welfare function for society.

We do not expect businessmen to devote a great deal of time and attention to maximizing the public interest. We assume that, although they will of course make some sacrifices to help the poor and advance the public welfare, basically they are concerned with benefiting themselves. Traditionally economists did not take the same attitude towards government officials, but public choice theory does. To simplify the matter, the voter is thought of as a customer and the politician as a businessman/entrepreneur. The bureaucracy of General Motors is thought to be attempting to design and sell reasonably good cars because that is how promotions and pay rises are secured. Similarly, we assume that the government bureaucracy will be attempting mainly to produce policies which in the views of their superiors are good because that is how their promotions and pay rises are secured.

In all these cases, of course, the individual probably has at least some willingness to sacrifice for the public good. Businessmen contribute both time and money to worthy causes and politicians on occasion vote for things that they think are right rather than things which will help them get re-elected. In both cases, however, this is a relatively minor activity compared to maximizing one's own wellbeing.

The only surprising thing about the above propositions is that they have not traditionally been orthodox either in economics or political science. Writers who did hold them, like Machiavelli in parts of *The Prince*, were regarded as morally suspect and tended to be held up as bad examples rather than as profound analysts.

Public Choice changes this, but even more important, by using a model in which voters, politicians and bureaucrats are assumed to be mainly self-interested, it became possible to employ tools of analysis that are derived from economic methodology.

As a result, fairly rigorous models have been developed which can be tested with the same kind of statistical procedures that are used in economics, although their data are drawn from the political sphere. The result is a new theory of politics which is more rigorous, more realistic, and better tested than the older orthodoxy.

While the basic thrust of the Public Choice work has been positive (directed towards understanding politics), from the very beginning it has also had a strong normative component. Students of Public Choice might modify Marx to read that 'the problem is to understand the world so that we can improve it'. Thus the design of improved governmental methods based on the positive information about how governments actually function has been an important part of Public Choice work, and is usually referred to as the theory of constitutions.

Before discussing this, it is necessary to outline briefly related discoveries in four general areas, viz: voters, politicians, the voting process which relates voters to politicians, and the theory of bureaucracy.

We begin with voters. One of the earliest discoveries of the new Public Choice (see Downs 1957, pp. 207–78) was that a rational voter would not bother to be very well informed about the votes that he cast. The reason is simply that the effect of his vote on his well-being is trivially small (see Tullock 1967a, pp. 100–14). Apparently voters have always known this, since empirical studies of voter knowledge show them extremely ignorant, but it was something of a revelation to traditional professors of Political Science. Further, this general ignorance of the voter is not symmetrical. The voter is likely to know a good deal about any special interest which he has. Further, organized special interest groups will put effort into propagandizing the voter in such areas. Thus the voter is not only badly informed, but what information he has tends to be biased very heavily in the direction of his own occupation or avocation. The farmer is much more likely to know the views of the candidates on farm programmes than their views on nuclear war. It could be said that even on the farm programme he is probably not very well informed, just better informed.

One should not exaggerate of course. The voter, simply by living and following current events in newspapers and on television, does acquire a certain amount of general information about politics. Not much of it seems to stick, however, and in any event it is very heavily affected by temporary fads. It should also be emphasized that some kinds of special interests of the voter are not in any real sense selfish. For example, in the USA many people are influenced in their vote by such institutions as Common Cause and Liberty Lobby and make voluntary cash contributions to them. Clearly, this is an expression by those people of their interest in good government, even though the two groups define this in a radically different way. There is no doubt, however, that a well organized special interest is apt to have more impact on any specific issue than either the general media or so-called public interest groups like Common Cause or Liberty Lobby, even though in the very long run, considering what one might call the 'general mystique' of government, the media are very important.

Consider next the politician. A politician is a person who makes a living by being elected by voters of the kind described above. Further, many politicians are themselves voters as, let us say, members of the House of Representatives. While in the latter capacity, although it is not true that politicians' information is as bad as that of the voter, a similar effect is still at work. An individual member of the House of Representatives or the House of Commons who switches one hour a week from general study of the issues on which he must vote to constituency service will normally reduce only trivially the quality of the legislation as it affects his constituency. On the other hand, by so re-allocating his time, he may materially improve his relations with his electors. Thus we would expect that politicians will be less well-informed on general matters than we would like.

This is simply one example of a large number of cases in which politicians' behaviour is not necessarily that which maximizes the public welfare: they vote in Congress and seek public positions in terms of what they think the voters *will* reward, not in terms of what they think the voters *should* reward. Since a politician knows that his constituents are badly informed, these two positions can be radically different. Nevertheless, if we are believers in democracy, which literally means popular rule, then the government should do what the people want and not what some wiser person feels that they should want. In any event, 'in order to be a great Senator, one must first of all be a Senator'.

Obviously the cost to the public of this kind of behaviour is quite considerable. It is particularly so when we think of the investment of resources and influence in the government which are, to a considerable extent, wasted. However, if we contrast functioning democracies with the other types of government which we observe, we are not likely to feel that democracies are markedly less efficient.

We now turn to the voting process, which connects the public to the politicians and the latter to the actual policy outcomes. Uninformed people think that this is basically a trivial problem, you simply count the votes. Unfortunately, this does not follow, even though the author of this essay is one of the few Public Choice theorists who regards the problems to be discussed next as being possibly illusory.

Condorcet, Borda, Laplace and Lewis Carroll and, in the 20th century, mathematical economists like Black and Kenneth Arrow discovered a set of mathematical problems sufficiently difficult to be taken as proof that democracy is either an illusion or a fraud. Basically, if we assume that all individuals can order various policy proposals, producing a personal ranking from top to bottom (indifference between alternatives being permitted) and that these orderings differ from person to person (and do not fall into a set of narrowly specified and rather unlikely patterns), then one of the following three phenomena can occur under any conceivable system of voting:

1. Endless cycling with A beating B and B beating C then C beating A.
2. An outcome which is dependent on the order in which the various proposals are voted on. (It should be pointed out in this connection that if this is so, and the people are well informed, voting on the order of voting reproduces the same problem.)
3. A situation in which the choice between alternative A and alternative B depends on whether alternative C (which in itself has no chance of winning) is or is not entered into the voting process. Most legislatures follow procedures which fall under the second of these possibilities.

If there is a possibility of arranging all of the alternatives in a single dimension with individuals having an optimal point and their preferences falling away monotonically as one moves away from that optimal point in either direction (single peakedness), then the problem is avoided. Unfortunately, most choices involve policies that differ from each other in more than one dimension and so cannot be arrayed in such a onedimensional continuum. Furthermore, voting on them one aspect at a time reintroduces the second of the problems above. Nevertheless, the assumption of single peaks (whose validity is probably due to voter ignorance) has been successfully used in much empirical work.

While there is no doubt about the mathematical accuracy of the proofs of the above propositions, the real problem is whether they are of great practical significance in voting. Unfortunately, this turns out to be an extremely difficult question whose solution is unlikely to be found in the near future. In essence there are two possibilities when we observe such voting bodies as the House of Representatives and look at the outcome. The first is that the outcome is essentially random, that is, matters are taken up in some order, that order determines the voting outcome and the members of the House do not realize that they could then change that outcome by changing the order in which the propositions are voted on. This possibility would imply that luck plays an immense role in democracy.

The alternative is to say that the outcome is manipulated by somebody who understands the situation and who has control over the agenda. The House majority leader, or the chairman of the Rules committee, is sometimes suggested as that person. This implies that we really have a dictatorship, one that is well concealed.

In my opinion, the indeterminacy thrown into the outcome by these propositions of social choice theory is actually quite small in practical terms. Thus the Chairman of the House Rules Committee may be able to change an appropriation bill by, say, one hundred thousand dollars, but not by an amount which (given the size of these appropriations) is particularly relevant (see Tullock 1967b). Among Public Choice theorists mine is a minority point of view. The majority, although it is deeply concerned about these problems, tends to ignore the implications of its point of view on the desirability of democracy as a form of government.

Empirical evidence has clearly demonstrated that agenda control can to some extent affect the outcome. This of course is going to surprise nobody. One does not need the complex mathematics of voting in order to realize that those members of any assembly who are in a position to control the order upon which things are voted have power. Similarly the control of what propositions are actually put before the voters can have considerable impact on the outcome. The demonstration of the empirical impact from agenda control, however, does not really support the theorems given above. Of course, we cannot say that the failure to find clearcut proofs that the outcome in a democracy is essentially either random or fraudulent (as would be implied by the mathematical work on voting) proves that it is not. The problem is difficult and subtle and in the present state of our knowledge must be left for further research. Meanwhile, we all go on with faith that the voting process produces an acceptable outcome even though mathematical investigation raises grave doubts.

Turning now to the theory of bureaucracy, once again Public Choice thought has worked a revolution. The traditional view was either that bureaucrats followed the orders of their political superiors or alternatively that they simply did what was right. Public Choice theorists, following the work of Tullock (1965), Downs (1967) and Niskanen (1971), believe that these are not proper statements about the bureaucrats' motives, although to some extent the bureaucrats do attempt to do what is right – including obedience to the views of their superiors. However, in modern societies where civil service legislation makes it all but impossible for the superiors either to dismiss them or even to reduce their salaries, the degree to which the bureaucrats are so compelled is moderate. Furthermore, in most civil service situations the power of a political appointee to reward his inferiors by promotion is very much restricted. Promotion decisions are to a considerable extent controlled by both legal and public-relations considerations which may compel a superior to promote someone whom he actually thinks has been sabotaging his policy.

While this is a characteristic of most modern civil service structures, there is no law of nature which says that government should be organized in this way. Traditionally, higher officials have been free to promote, demote or dismiss this subordinates. Even here, however, the fact that the higher official cannot possibly know everything that is going on at the lower ranks means that his control gradually diminishes as one moves away from his position down the pyramid of ranks.

For example, in the USA it was recently discovered that it is not possible for the Secretary of Defense to know the specifications which a civil servant, located at a vast distance down the pyramid, produced for a new coffee pot for military aircraft. In this case, the civil servant who specified a coffee pot capable of withstanding a crash that would kill the entire crew of the plane was neither dismissed nor even reprimanded. Indeed the newspapers that reported the story did not even mention his name, but instead concentrated on the Secretary of Defense. In 1870 a military procurement agent who make a mistake like this (and which got into the newspapers) would have found it necessary to hunt for a new job within an hour or so.

Basically the average employee in a bureaucracy is interested in retaining his job and gaining promotion and for this purpose wants to please his superiors. Under the old-fashioned system where he had little job security, and where promotion was determined strictly by his superiors, there was considerable pressure on him. In present circumstances, where to all intents and purposes he cannot be dismissed and where even his promotion is to some extent protected from political intervention by his superiors, this pressure is less important. However, even in a different case, in which he did indeed want to please his superiors, this would not necessarily lead to activity which is in the public interest. That would depend on the political situation of the party or individual who at that time was in control of his branch of the government.

This attenuation of control, in which much of what is done by lower-ranking officials is simply unknown to those of higher rank, is characteristic of all bureaucracies. There are however various ways by which the higher ranks can become, to some extent, aware of what is being done by the lower ranks. Undoubtedly the most efficient of these is simply an accounting system. In the case of a private company, whose motive is making money, the accounts do a reasonably good job (no more) of signalling what the various lower ranking officials are contributing to that goal. When we turn to government, however, we have the combination of a set of objectives that are either vague or not clearly specified, and a situation where there is no accurate way of measuring the contribution of each person to those objectives. Under such circumstances, control is much more severely attenuated.

When we have a civil service structure which separates the individual from much of the control power of his superiors, the problem is even more severe. Whether an individual bureaucrat works hard or not, prepares himself or herself well or not, is largely a matter of individual choice. As a rough rule of thumb, those people who do work hard and prepare themselves well are those people who have their own idea of what government should do in their particular division and work hard at that. In a way they are hobbyists. It should be said however, that their hobby is normally motivated by a desire on their part to maximize what they think is the public good. In other words, they are usually well-intentioned individuals who can be criticized only in that their idea of the public good may or may not coincide with that of their superiors. If it does not coincide, this does not prove that they are wrong and the superiors right, but it does mean that the government is not apt to follow a coordinated policy. In times past, it used to be normal to refer to the US Department of State as ‘a loose confederation of tribal chieftains’. The phrase is not used any more, but as far as I can see this is only because the confederation itself has broken down.

Bureaucrats normally have several private motives. One is, of course, simply not to work too hard – a motive which does not seriously affect the hobbyist described above. Another is to expand the size of one’s own department and in the process of so doing, being willing to go along with the expansion of all the rest. A third is to improve the ‘perks’ that accompany the particular position (see Migue and Balageur 1974).

Note that this is not intended as criticism of the bureaucrat. We would expect anyone who is given the kind of opportunities that are given to bureaucrats to do more or less what they do. However, the consequence is that large bureaucracies tend to grow larger, tend as they grow larger to follow less in the way of integrated policies and more in the way of policies that develop in the lower reaches of the pyramid, and tend in fact not to work terribly hard (see Bennett and Orzechowski 1983).

The problem is multiplied when bureaucracies become very large, because the members of the bureaucracy can vote. Furthermore, empirical evidence (see Bennett and Orzechowski 1983) shows they vote more frequently than non-bureaucrats. Thus their percentage in the voting population is somewhat larger than their percentage in the actual population (see Frey and Pommernhe 1982). Thus, the political superior must consider the people working for him as in part his employers rather than his employees. He may not be able to fire them, but in the mass they can fire him. Altogether, the system is not well designed and does not work very well.

So far we have been talking about Public Choice and what has been learned, but not of the lessons of a normative nature that have been drawn i.e. the theory of constitutions. It is to this that I now turn.

Not all students of Public Choice favour the same reforms in each area. Further, some have not specifically said what reforms they would prefer because they believe that not enough is yet known about the process to be able to suggest improvements. Nevertheless, there are several rather general propositions which most students would agree upon as ways of improving the functioning of government. In a discussion as brief as this, it is not possible to include all the differences of opinion and all the modifying clauses which would be appended to each suggestion for reform. Thus the reader should not assume that everyone studying Public Choice agrees with all the propositions which follow.

To begin with the voter, no student of the subject has any idea of how to improve the voters' information. With respect to voting itself there have been some proposals for improved voting methods, but no widespread support exists for any particular improvement. In spite of this, I think it can be said fairly that most students would like to see voters vote more than they do now, favouring more direct voting on issues, and legislatures with larger membership (so that the connection of an individual voter and his representative is closer).

The basic desire to give voters more control of the mechanism is not based on any false idea of how well the voters are informed. It is simply that the voters are the only people in the whole process who do not have an element of systematic bias in their decision process. They may be badly informed, but what they want is their own well-being. The well-being of its citizens should be the objective of the state. When we turn to other parts of the government invariably we find at least some conflict between the interests of the officials and the interests of the average man. Thus increasing the average man's control is not particularly likely to improve the efficiency of the government using some abstract definition of efficiency. But it is likely to make the government more in accord with the preferences of the common man; i.e. it brings us a little closer to the objective of popular rule which is supposed to be what democracy is about. Those who do not favour popular rule would not regard this as desirable, but there are few elitists among the students of Public Choice.

The actual decision-making procedures used in the legislatures have been widely discussed and some proposed improvements command wide acceptance. First, many would like to have at least one house of the legislature elected by proportional representation. Secondly, Buchanan and Tullock's arguments in *The Calculus of Consent* (1962) for bicameral legislatures have generally been accepted. The further suggestion there that more than a simple majority is desirable for most legislation is seldom directly criticized, but is not so widely approved. The argument that this higher-than-majority requirement would change the structure of the logrolling process in a favourable way has seldom been directly criticized, but the asymmetrical effect of such a rule (i.e., the status quo is retained unless a reinforced majority can be obtained to change it) offends some people.

Turning to the bureaucracy, there is much more agreement on reform. First, that a bureaucracy should be brought more firmly under the control of the political leaders is, I think, uniformly accepted. The dangers of this are recognized – but there are various ways in which the higher officials could be given the right to discipline civil servants while still reducing their power to fill the government with their cousins.

Apart from such straightforward proposals for changes in the personnel structure there are other ways of putting pressure on the government. The first is to work some competition into the system. Currently, not only do most government departments have a monopoly over whatever function they perform, but almost every proposal to increase the efficiency of government takes the form of eliminating what little competition has popped up. Competition between government departments should be encouraged rather than discouraged.

Finally, it may be possible to 'contract out' government activities or literally transfer them wholely to the market. The mere threat of this will frequently lower the cost of government activity. Having several private companies bidding for a government service, however, is better.

It can be seen that at the concrete level, those who study Public Choice have been able to provide more in the way of suggestions for reform within the bureaucratic structure than in the higher level parts of democracy where the voters control the legislature, and the legislature and executive then control the bureaucracy. This is unfortunate but not surprising. Nevertheless, there are suggestions for improving the whole structure of government and with time, it is hoped, there will be both more ways of making improvements and better scientific evidence that the 'improvements' are indeed improvements.

Public Choice is a new and radical approach to government, but its firm foundations in economic methodology mean that we have more confidence in its accuracy than with most new ideas. Further, it has by now been empirically tested very thoroughly. Government is the solution to some problems and the source of others. Public Choice shows strong promise of being able to reduce significantly the difficulties we now have with democratic government.

See Also

▶ Black, Duncan (1908–1991)
▶ Constitutions, Economic Approach to
▶ Social Choice

Bibliography

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

Bennett, J.T., and W.P. Orzechowski. 1983. The voting behavior of bureaucrats: Some empirical evidence. *Public Choice* 41(2): 271–284.

Black, D. 1958. *The theory of committees and elections*. Cambridge: Cambridge University Press.

Buchanan, J., and G. Tullock. 1962. *The calculus of consent: Logical foundations of constitutional democracy*. Ann Arbor: University of Michigan Press.

Downs, A. 1957. *An economic analysis of democracy*. New York: Harper & Row.

Downs, A. 1967. *Inside bureaucracy*. Boston: Little, Brown.

Frey, B.S., and W.W. Pommernhe. 1982. How powerful are public bureaucrats as voters? *Public Choice* 38(3): 253–262.

Migue, J.L., and G. Balageur. 1974. Towards a general theory of managerial discretion. *Public Choice* 17: 27–43.

Niskanen, W. 1971. *Bureaucracy and representative government*. Chicago: Aldine-Atherton.

Tullock, G. 1965. *The politics of bureaucracy*. Washington, DC: Public Affairs Press.

Tullock, G. 1967a. *Toward a mathematics of politics*. Ann Arbor: University of Michigan Press.

Tullock, G. 1967b. The general irrelevance of the general impossibility theorem. *Quarterly Journal of Economics* 81: 256–270.

Tullock, G. 1981. Why so much stability. *Public Choice* 37(2): 189–202.

Public Debt

James M. Buchanan

Abstract

Classical principles of public debt limited debt financing to non-recurrent, extraordinary or temporary demands. Keynesian macroeconomics, viewing budget deficits as the only means of financing demand-increasing deficits during depressions, overlooked the exchange between government and lenders in debt-financed public expenditure. In the post-Keynesian 1970s and 1980s, governments explicitly used debt to finance ordinary public consumption, including transfers, which was equivalent to a destruction in national capital value and raised the prospect of default. Fiscal responsibility demands that the classical principles of public debt must eventually return to general acceptance.

Keywords

Assets and liabilities; Barro, R.; Buchanan, J. M.; Budget deficits; Capital value; Default; Fiscal responsibility; Keynesian revolution; Neutrality theorem; New classical macroeconomics; Public debt; Ricardian equivalence theorem; Ricardo, D.; Taxation

JEL Classifications

H6

Public debt (government debt) is a legal obligation on the part of a government to make interest and/or amortization payments to holders of designated claims in accordance with a defined temporal schedule. Public debt is created through government borrowing from individuals, corporations, institutions, and other governments. Borrowing is part of a bilateral exchange process in which lenders transfer funds to government and government, in turn, transfers to lenders designated instruments that represent claims on government revenues over a series of periods subsequent to that in which the borrowing occurs. In simple balance-sheet terms, public debt is a liability item on the government's account, and an asset on the combined accounts of the holders of the debt instruments.

In its essential respects, public debt is not different from the debt of individuals or non-governmental institutions. The positive analysis is equivalent over the several settings, and the normative principles for the use of debt are the same. These two propositions are not universally accepted by economists, whose understanding and analysis of established classical principles were eroded in the emergence of the macroeconomic mind-set of the postKeynesian era. Because of the continuing confusion and ambiguity in the basic economic theory of public debt, extended discussion is required on what should otherwise seem quite elementary analytics.

Taxes, Money and Public Debt

In order to finance spending programmes (including transfers) government must first secure revenues. Three means are available to governments with independent monetary systems: taxation, money issue, and public debt. Only two means are available to subordinate governments without money creation powers or to national governments that tie domestic currencies to external forces in the international economy: taxation and public debt.

It is useful to make two critical distinctions between taxation on the one hand, and public debt on the other. With taxation there is only one possible 'exchange' embodied in the combined fiscal process, that effectuated via government between individuals as taxpayers and individuals as beneficiaries of government spending programmes. This two-sided fiscal operation is 'exchange' only in the aggregative sense that persons in the community secure benefits from the taxes paid by persons, whether or not the taxpayer and beneficiary groups are overlapping. However, even if the fiscal process satisfies the aggregative and the individualistic efficiency standards such that all persons pay tax-prices, at the relevant margins, equivalent to public-goods benefits, coercion is required to implement the solution. The basic fiscal 'exchange' is not, and cannot be, voluntary.

With public debt issue, by comparison, *two* 'exchanges' are involved in the combined fiscal operation, one, the political 'exchange' analogous to that in taxation, and the other the whole set of privately negotiated and wholly voluntary exchanges between government and those who lend funds to government. Failure to recognize the double set of exchanges that public borrowing combined with spending of the proceeds embodies is the source of major confusion in determining the location of the burden of debt, to be discussed below.

A second critical distinction between taxation and debt issue lies in the temporal difference in the politically determined imputation of the liabilities that are made necessary by the fact of government spending. With taxation, these liabilities are imputed to those persons and institutions that make the funds available directly to government in the period of the spending operation. With public debt, by contrast, there is no current or initial-period imputation of fiscal liability. Revenues are secured from those who lend voluntarily, who do so in exchange for promises of future period interest and amortization payments, and not in 'political exchange' for the benefits of spending programmes, even indirectly. With government borrowing (debt issue) the ultimate fiscal liability made necessary by the spending programme in the initial period is postponed. This liability is placed, in the aggregate, on taxpayers in periods subsequent to that in which the debt is issued. The aggregate liability is not, however, imputed or assigned to individuals or to groups of individuals. The postponement of liability implies, therefore, postponement of payments for public programmes.

The Ricardian Theorem on the Equivalence Between Taxation and Government Borrowing

David Ricardo (1817, 1820) advanced a theorem to the effect that taxation and government borrowing are logically equivalent. This Ricardian theorem was rediscovered by macroeconomists in the 1970s, notably by Robert Barro, and it became an important element in the 'new classical macroeconomics' of that period.

On its face, the theorem seems to reject the critical distinctions between taxation and debt discussed above. In what respects are taxation and government borrowing alike rather than different? Both extract revenues from private citizens and transfer these to government for spending on public programmes. The basic Ricardian logic does not reject the difference in status between the taxpayer, who faces governmentally imposed coercive levies, and the bond purchaser, who voluntarily transfers funds to government in private exchange for future-period interest and amortization payments. The theorem does, however, reject the second distinction noted above, that which involves any postponement of the costs of public spending. In the Ricardian model, individuals recognize that any issue of debt by government embodies a commitment to meet payments in future periods. These payments can be converted into individualized shares in the aggregate liability, discounted, and capitalized into a present-value measure, which can then be reckoned as a liability item on individual initial-period balance sheets. If persons think that they will live forever, or if they have intergenerational bequest motives that cause them to act as if their lives are infinite, the liability items, summed over all persons, will just offset the value of the debt in the balance sheets of those who hold debt instruments.

If the equivalence theorem is valid, there are important macroeconomic consequences. If persons treat government debt equivalently with taxation in all respects other than the actual timing of the payments, they will make portfolio adjustments as required by this differential timing. When debt is issued, persons will, knowing that payments must be made in future periods rather than currently, put aside some funds to facilitate such payments. There will be no difference between tax and debt financing in their effect on consumption and investment spending. Individuals, as taxpayers-citizens, will when debt is issued, increase savings to allow a share of the future-period debt obligations to be met. However, this increase in saving will equal the full value of the debt only if the governmental outlays take the form of ideally efficient transfers. This *neutrality theorem,* which we may associate with

Barro, is more restrictive than the Ricardo theorem. If the governmental outlays are made for the provision of real goods and services, the neutrality theorem may not hold although debt financing and tax financing of these outlays may still exert equivalent effects (the Ricardo theorem).

The Ricardo equivalence theorem (along with the stronger neutrality theorem) is best evaluated as an extreme model of rational individual behaviour under idealized sets of circumstances. Ricardo himself recognized that persons did not, in fact, treat taxation and debt in the same way. All persons do not act as if they live forever; persons differ in age as well as interest in future-period tax obligations. Further, taxes are not lump sum, interpersonally or intertemporally. More importantly, there is no assignment of present-value liability among persons such as would allow portfolio adjustments to be made in the manner postulated in the equivalence theorem setting. Even if individuals do recognize that public debt does embody future-period tax liabilities, they cannot reduce this aggregate to individualized shares in any plausible reckoning.

The central flaw in the equivalence theorem stems from the logic of debt itself, which may be illustrated by analogy with private behaviour. Why does a person borrow? He does so in order to rearrange spending temporally. If borrowing and current payment (the private analogue to taxation) are equivalent, there is no point in the exercise. As an institution, borrowing has as its purpose the adjustment of spending flows through time. Governments, as agents are citizens, borrow for analogous reasons. There is no *raison d'être* for public debt if this instrument is behaviourally equivalent to taxation.

The empirical evidence gained from straightforward observation of modern politics points clearly toward rejection of the equivalence theorem. Taxation and debt are not treated as identical by voting constituents, as is suggested by the fact that politicians responsive to constituents are not indifferent as to the mix between these instruments. Within broad threshold limits, debt financed outlay arouses less political opposition than tax-financed outlay of comparable magnitude. The observed US Federal deficits of the 1980s could not have been eliminated by tax increases without generating significant political struggle.

Classical Principles of Public Debt

Both the positive analysis of and the normative precepts for public debt were broadly understood by the classical economists, and these principles were carefully articulated in the dominant theory of public debt developed in the 19th century. There is no essential difference between the government account and the account of an individual or private firm in the classical model. Borrowing is a means of raising revenues that allows the borrower to put off or to postpone payments. It is a means of adjusting spending needs to revenue flows over time; in effect, borrowing allows intertemporal trades to be made.

For the government, as for the individual or firm, there would be no basis for borrowing unless the burden of payment could be delayed in time. By the very meaning of debt, therefore, there must be a shifting forward of burden intertemporally. The ultimate payments for the enhanced spending programme during the initial period when debt is issued must be borne exclusively in later periods.

From this straightforward and indeed simple analysis, normative principles for public debt creation emerge. Resort to debt financing is indicated only with nonrecurrent or extraordinary demands, or requirements for public spending that are expected to be temporary. Traditionally, such demands were associated with war emergencies, and the principles of fiscal prudence dictated that debts accumulated during was periods would be retired when the emergency spending demands were past. In addition to these extraordinary spending justifications for resort to borrowing, government is also within bounds under classical norms when debt is issued to finance genuinely productive capital projects, analogously with private firms making capital investments.

When capital spending is debt financed by government, the principles suggested that a scheme for debt retirement be put in place to insure that the pay-off period corresponds to the income-yielding period from the investment asset.

Public Debt in Keynesian Macroeconomics

The classical principles of public debt were not understood by the pre-classical mercantilists and these principles were also questioned by a few fiscal and monetary expansionists in the pre-Keynesian period. Only with the 'Keynesian revolution' in economic thinking in the middle of the 20th century, however, did a rejection of these classical principles become a 'new orthodoxy'. An analysis of public debt, along with relevant normative implications, came to be dominant during the 1940s and 1950s that sought to contradict basic elements of the classical model.

As noted earlier, the classical principles are starkly simple, and are based on the essential similarity between the government and the individual account. The Keynesian logic rejected this analogy. Specifically, the argument denied that public debt embodies any shift of burden onto taxpayers in periods of time subsequent to debt issue, a temporal shift that was acknowledged to occur in both private debt and external public debt.

The conclusion that public debt could involve no intertemporal shift of burden emerged from an undue concentration on the macroeconomic aggregates and an overlooking of individual adjustments to macroeconomic instruments. Again, the logic is seemingly quite straightforward. Resources are used up only in periods when spending programmes take place; if government borrows to finance spending on guns, the resources that go into producing these guns must be given up by some persons during the period and not later. With internal public debt, therefore, the burden of war spending could not, by definition, be transferred forward.

The basic flaw in the argument is clear from the discussion in section I. The two-part exchange that debt-financed public spending embodies is overlooked. The argument fails to recognize that those who actually give up claims over resources during the period of the spending do so because they voluntarily exchange funds for promises of interest in future periods. These purchasers of bonds do not, in any sense, 'pay for' the benefits of the public spending programme. The fact that these persons may also be members of the community of citizens-taxpayers is irrelevant for the temporal location of burden. Taxpayers in later periods are faced with claims against their incomes that must be met, and which exist only because of the initial-period debt issue. If, indeed, the Keynesian orthodoxy of public debt were valid, economists and finance ministers would have discovered the fiscal equivalent of the perpetual motion machine. No non-voluntary transfers of revenues are required to finance spending in the initial period, and, if there is no burden on future-period taxpayers, the spending that is carried out would have required no burden on anyone.

The Keynesian argument was driven by a stance on policy that viewed public debt as the only means of financing demand-increasing deficits during periods of depression. The primary policy instrument of Keynesian economic policy was the budget deficit, and there was an elementary failure on the part of pro-Keynesian economists to recognize that demand-enhancing deficits could be financed with noninterest bearing money creation. If this macroeconomic objective is the only justification for the creation of budgetary deficits, it becomes totally unnecessary to impose the future-period taxes that debt interest reflects. Money creation in such settings carries with it no future-period burden.

Public Debt and Deficits in PostKeynesian Politics

The Keynesian replacement of classical principles of public debt was never total, and in the late 1950s and early 1960s there was a reemergence of economists' support for the earlier analysis, along with its normative implications. There was not, however, a 'paradigm shift' at all comparable to the earlier overthrow of the classical analysis.

Somewhat begrudgingly perhaps, economists of the 1960s came to recognize the deficiency in the simplistic Keynesian logic, but there was no general reaffirmation of classical principle. The discussion of public debt that characterized the 1960s, 1970s and 1980s remained confused by an admixture of the two contradictory models of analysis. Economists seemed to concentrate attention on the secondary and tertiary macroeconomic consequences of debt-financed deficits; their considerable formal skills were directed toward attempts to extend the ancient Ricardian theorem.

While confusion and ambiguity described economists' discussion of public debt, the politicians had learned the Keynesian policy lessons with roughly a two-decade lag. By the early 1960s, the 'old-time fiscal religion', based on adherence to the normative precepts of the classical analysis, had lost its constraining influence. The political leaders of the 1960s and beyond had learned that demand-enhancing deficits may be justified in some economic settings. Their natural proclivities to spend without the levy of taxes on constituents caused them to look on economic settings in a biased or one-sided fashion. The idealized Keynesian policy set – deficits in depression, surpluses in booms – proved to be unworkable in democratic politics.

The regime of apparently permanent debt-financed deficit spending was born. During the 1970s and 1980s, for the first time in modern fiscal history, governments explicitly used debt to finance ordinary public consumption outlay, including transfers. This fiscal operation, considered in isolation, is equivalent to a destruction in national capital value. When persons, privately or publicly, abstain from consuming current income, capital value is created. When persons, privately or publicly, consume more than current-period income, capital value (defined as the discounted present value of anticipated future incomes) is destroyed. For both individuals and governments, resort to borrowing allows a 'using up' of future-period income as a means of increasing current consumption, just as resort to saving allows a 'using up' of current income (in an opportunity cost sense) to increase future-period consumption. Only if borrowing is used to finance genuine capital investment, private or public, will the net effects be intertemporally neutral; only in this case will the capital value of anticipated income be unchanged. With debt-financed public consumption, the present value of anticipated future incomes of persons in the polity is reduced relative to that which might have been maintained in the absence of the combined fiscal operation. The fact that some or all of the debt is held by foreigners rather than citizens does not modify this conclusion.

Consider the case where debt instruments are purchased exclusively by domestic citizens, who may also be future period taxpayers. Securities are purchased voluntarily; hence, there is no change on the asset side of purchasers' balance sheets at the time of debt issue. Purchasers could, alternatively, hold or buy privately issued securities with equivalent yields. On the other hand, the fiscal operation does place debt-interest claims against the anticipated private income flows of citizens as taxpayers. There is a net increase in the present value of liabilities on properly calculated individual balance sheets. This increase in the value of liabilities is, of course, equivalent to the value of the debt instruments. But these two items are *not* offsetting since there are two fully offsetting items on the asset side of the accounts, leaving no net change from this side. The combined fiscal operation necessarily reduces net worth, or capital value, in the economy, so long as the government outlay does not generate anticipated income flows from public assets, a result that is ruled out with pure public consumption.

Return to Classical Principles?

Public debt is a topic in political economy in which the level of understanding experienced serious retrogression over the course of the middle decades of the 20th century. Policy-motivated macroeconomic confusion generated political spillovers that remained in the 1980s. Economists seemed unable to contribute to clarification in analysis, in part because they were reluctant to drop either suprarational models of individual behaviour or erratic manipulation of data. The classical analysis, out of which emerged precepts that offered simple guidelines for governmental fiscal authorities, no longer commanded widespread support, either among political economists or among politicians and, indirectly, their constituencies. Governments in the 1980s were observed to be financing sizeable shares of their public consumption outlays by interest-bearing debt. Interest outlays made up ever-increasing proportions of total government budgets.

The simple logic of compound interest guaranteed that the budgetary regimes observed in the 1980s were not sustainable. Default on government's debt obligations becomes increasingly attractive to politicians as interest charges mount and as borrowing rates for new issues of debt simultaneously increase. Default on public debt has occurred often in history, both through explicit destruction of real value obligations and by means of inflation.

The ultimate prospects for default may be generally recognized, but the political difficulties in restoring some adherence to classical norms cannot be overlooked. Once debt-financed deficit spending for public consumption came to be an element in the quasi-permanent *status quo,* attempts to restore budgetary balance faced enormous political opposition, as indeed the events of the 1980s demonstrated. To reduce the deficits, and hence merely to reduce the rate of increase in public debt issue, governments must resort to tax increases or to spending cuts, both of which arouse political opposition. Those taxpayers who must bear the burden of continuing debt are, at best, only partially and indirectly represented in the decision structure of democratic politics.

Restoration of the classical principles of public debt seemed unlikely from the temporal perspective of the 1980s. The old-time or pre-Keynesian 'fiscal religion' did exert an influence on the behaviour of politicians, and through them, on governments. Public debt, as a revenue-raising instrument, has an appropriate and well-defined use as a means of allowing governments to alter the time stream of payments for extraordinary outlays. There was nothing comparable to a 'fiscal religion' in postKeynesian politics. Public debt, as it was actually used in the years after the 1960s, became a mere balancing element between proclivities of politicians to tax on the one hand and spend on the other. The absence of immediate fiscal breakdown was explained by some residual carry over of classical norms, some introduction of a Ricardian-like consciousness of future tax liabilities, and some fear of default risk on the part of prospective lenders. But the situation observed over the decades of the 1970s and 1980s could not have represented temporal stability.

The classical principles of public debt, whether they be labelled as such, must eventually return to general acceptance if the fiscal responsibility of governments is to be maintained. Whether or not this acceptance comes before or after a sequence of default-engendered fiscal crises could not be predicted from the temporal perspective of the middle 1980s.

See Also

- ▶ Government Budget Constraint
- ▶ Public Finance
- ▶ Ricardian Equivalence Theorem

Bibliography

Adams, H.C. 1893. *Public debts*. New York: Appleton.

Barro, R.J. 1974. Are government bonds net wealth? *Journal of Political Economy* 82: 1095–1117.

Bastable, C.F. 1895. *Public finance*. 2nd ed. London: Macmillan.

Buchanan, J.M. 1958. *Public principles of public debt*. Homewood: Richard D. Irwin.

Buchanan, J.M., and R.E. Wagner. 1977. *Democracy in deficit*. New York: Academic.

Ferguson, J.M., ed. 1964. *Public debt and future generations*. Chapel Hill: University of North Carolina Press.

Harris, S.E. 1947. *The national debt and the new economics*. New York: McGrawHill.

Lerner, A.P. 1948. The burden of the national debt. In *Income, employment, and public policy: Essays in honor of Alvin H*, 255–275. Hansen: Norton.

Leroy-Beaulieu, P. 1912. *Traité de la science des finances. Vol. 2: Le budget et le crédit public*. 8th ed. Paris: Alcan.

Ricardo, D. 1817, 1820. *Works and correspondence,* ed. P. Sraffa, vol. 1, vol. 4. Cambridge: Cambridge University Press for the Royal Economic Society, 1951.

Public Economics

Serge-Christophe Kolm

The Modern Economic Theory of Public Actions: Aim and Scope

Public Economics is the study of the public economy, i.e., of economic questions which are not purely market, intra-household or intra-firm, with emphasis on logic-intensive (scientific) analysis and on ethical–normative questions.

'What must the government do?' is therefore a large part of the problem to which Public Economics sets out to provide answers, with the widest understanding of the 'government', and no real limitation in scope as to which acts it applies – this is because the term 'economic' is so extensive and in studies of this kind often defines more an approach, a point of view, than a domain in the substantive sense. Public expenditures, taxes, regulations of many kinds, public production and prices, public debt and money, exchange rate policies, etc., are the variables to be chosen. With a public sector using 30 to 70% of GNP in 'Western' developed countries – and much more in 'Eastern' socialist ones – with multifarious regulations, a vast array of bodies in between the purely public and the purely private, plus the global macroeconomic regulation of the market sector, the scope for Public Economics is a priori vast. But where the public sector must, or must not, lay its hands is the first question of Public Economics (not to intervence is a possible solution to the problem of what to do or how to do it). The discipline aims to provide specific and scientific answers to this question, rather than leaving it to ideologies. In fact, most Public Economics has been concerned with the interface between the private and the public sectors, seeking to analyse its place and structure and advise about it.

Institutionally, Public Economics is concerned with the behaviour and existence of public bodies in the largest sense – governments at all levels, quasi-public organizations such as public social insurance or health and education services in many countries, public firms and 'utilities', etc. It has a deep interest in the border or common domain between the public economy and the market, that is the voluntary associations of various kinds and aims. And, of course, the factual working of the private economy (markets and intra-agent), and normative judgements about it, are crucial for Public Economics, since it undertakes to advise on how to correct its defects.

Public economics, in its final aim, is thus largely an applied normative intellectual endeavour, which finds both means and aims in man and society and their environment, and for which constraints and possibilities contain capabilities and motivations both of private agents and of persons or groups constituting the public sector (civil servants, politicians in government or in parliaments, etc.). However, at some point social ethics must intervene in the reasons for public sector activities in a way in which it does not for those of the private sector – be it through an ethic of the Public Service, a constitution, the higher aims of politicians or the social objectives of the voting electorate – if only because the State has the monopoly of dominant force.

In societies where the economy is based primarily on a market system supported by an ethos praising its virtues, Public Economics naturally began with analyses of what came to be called 'failures' of the market, i.e., aspects where some intervention or another kind of organization and decision-making might improve performance. Phenomena such as public goods, externalities, non-discriminating monopolies, absence of futures markets, barriers to entry, poverty and distributive justice, widespread unemployment, inflation, destabilizing speculations, process preferences on the type of economic relations, effects of exchange and market relations on man and on culture and society, norm and status roles of prices and wages, many kinds of ignorance, etc., were in turn analysed with a view to helping, supporting, supplementing, correcting or straightening the market, or eventually replacing some of its elements by other defined decision-making processes or critieria. Public Economics was then supplemented by the analysis of the possibilities and limitations of non-market decision-making, in particular of political and public administrative choices – a research programme referred to as 'Public Choice'.

Born in the early 19th century and enriched by a small but steady flow of studies, Public Economics – apart from important but ill-found macroeconomics – experienced a burst of activities after the mid-1960s (when it was named), a time of rapid government growth in the West. This development was the Old Public Economics, which was superseded in the 1980s by the New Public Economics. Both engage in refined analysis of economic relationships. But the ethical part of the former was either primitive or incomplete (Pareto-optimality) or unacceptable (Social Welfare Function); in fact it was not even explicitly thought to be ethics, and this seriously impaired the practical acceptance, application and usefulness of its ideas. By contrast, the New Public Economics takes into account the relevant properties of the ethical question (a development that was influenced by the simultaneous upsurge in the analysis of social ethics in political philosophy, which itself borrowed much from economics).

[The term Public Economics, like its original ideas, seems to have been coined in France, a market economy endowed by history with a large public sector, important public firms, a civil service ethic and institutions, an ideology for some economic role for the State (from Colbert to the Plan), and bodies of both public-minded and mathematically oriented State engineers. The present author's *Fondements de l'Economie Publique, introdùction à la théorie du role économique de l'Etat* (1964) probably first used the term in print, and Leif Johansen's *Public Economics* (1965) soon followed. But, of course, Richard Musgrave's previous *Theory of Public Finance* (1959) was Public Economics, and a crucial landmark in the field. From 1966 on there came, under this name, regular meetings, an association and a journal founded by A. B. Atkinson.]

The Ethical Foundation

For Public Economics to reach its results, the problems it meets must be evaluated by ethical criteria, which are thus as important as technical and behavioural structures. Society reveals these criteria in the form of desires and rights. The traditional ideas supplied by) 'Welfare Economics' are of little help. The maximization of a Social Welfare Function depending upon individuals' preferences is in fact an ethic rejected by everybody in society because the resulting allocation concerning one person would depend upon structures of preferences of other persons in ways which are considered unacceptable. Pareto-optimality (the impossibility of improving the situation for one person without hurting some other) does not provide the unique solution and distribution which is required for practical application, and the corresponding 'efficiency' alone is of no avail – that is why so many proposals by economists are rejected by the practical men who would have to apply them. Various 'compensation schemes' have the same defects. Furthermore, anything the public sector does is bound to displease someone, so that with respect to these relevant variables the existing situation is always Pareto-optimal, and the practical problem is to choose between various solutions all having this property. Yet Pareto-optimality is a useful subordinate and partial criterion as will be seen below.

De facto, the most pervasive form of social judgement is that such specific action is 'legitimate' because it respects legitimately held or acquired rights ('legitimate', an ethical notion, is distinct from the juridical concept of 'legal'). Then, deducing from individual freedom both the legitimate ownership of oneself and thus of one's own labour, and the right to give or give up services or rights (the property of an object is a bundle of rights concerning its use) – hence the right to freely (unanimously) exchange or agree between two persons or more – there results, if previous relevant acts were legitimate, the basic legitimacy of the free market, of voluntary associations, and of the resulting allocations of rights and ownerships.

But legitimate, free and unanimous exchanges or agreements between any number of persons may fail to be directly and explicitly achieved because of a series of phenomena which are difficulties, costs or impossibilities in the realization of transactions, in the informational requirements of exchange, of bargaining, of transmitting demands and propositions (and to begin with of knowing the other parties), in the writing of contracts and in guaranteeing their implementation (through checking facts and acts and enforcement), in exclusion from the benefits of a public good or of receiving a positive externality or of creating a negative one, in relations between persons who do not exist at the same time, etc.

To implement by public means what these 'impeded' legitimate agreements *would* have decided if they had spontaneously emerged is the unavoidable ethical basis of modern Public Economics: it is unanimously desired at levels of both specific decisions and of general freedombased principles. This precludes neither a primary concern for welfare and redistribution, nor the necessity of the use of force in the implementation. On the contrary, Public Economics is concerned with welfare and redistribution because people are, and one of its main tasks is to find out what transfers direct general agreements would have produced. Secondly, private contracts are made of two parts: the voluntary unanimous agreement between the parties, and the implementation which is obligatory under the threat of public force. In the implicit contracts which establish this public ethic, the first part does not show explicitly, and there remains only the second one, compulsory implementation: hence the apparent pure obligation in public actions such as taxation or regulation.

In order to find out (or to make the public sector achieve) the content of these implicit, putative and tacit contracts, modern Public Economics draws upon theories of choice, exchange and bargaining, statistics, polls and observation of political processes, and it offers advice about those political structures or processes which would help to reveal the necessary information or achieve the right outcome (aspects of constitutions, referenda on public issues and their financing, optimal decentralization of public choices and 'levels of government', etc.). Each of these implicit contracts (called) 'liberal social contracts') determines a set of public actions and expenditures, and of taxes, which is desired by a group of citizens and violates no legitimate right of any citizen. It is such that no sub-group of the concerned persons could quit this cooperative agreement while securing to each of its members a situation he prefers, given that this person is in general influenced by what all others find their best interest to do (this last proviso differentiates this concept from the traditional one called 'the core') and if this sub-group is not itself impeded in a similar manner, etc. This condition implies Pareto-optimality with respect to the constraints which have not been abstracted from in defining the putative free agreement (the omitted constraints bear upon information, transaction, possibilities to constrain or exclude people, etc.).

We shall consider now the various elements that constitute Public Economics.

Collective Concerns and Consumption, Public Goods

The existence and analysis of collective concerns and consumption, and 'public goods', is central in Public Economics.

A 'collective concern' is something which concerns several persons or agents. Something which pleases several persons together came to be called a 'public good' for them. If it displeases them it is a 'public bad', the decrease of which is a public good for them. In particular this good can be a genuine good or service in the sense of economics, and then there is a collective consumption (either final or intermediary or both together). 'Non-rivalry' in consumption is the term sometimes used (Musgrave) to characterize the difference with 'private goods'.

A private profit-making agent can produce a public good if he can exclude the would-be beneficiaries from its benefit, so as to grant them access only in exchange for payment of a fee or price or toll. These receipts may cover the costs of producing the good and allow a profit. The producer's problem is to obtain information about the willingness to pay, since if he asks too little he may not be able to cover costs, and if he asks too much for some access he may prevent it, whereas the user would be ready to pay more than the individual specific costs he causes (which may be zero). In addition, the producer may not be able to discriminate his price sufficiently between the beneficiaries, not only because of difficulties in discerning differences (which is an information problem), but also because he may not be able to prevent them from reselling between themselves the rights to access or private goods produced thanks to the public good, or for reasons of practicability or fairness. These problems of course interfere with the quantity of the good the producer decides to create. The lack of unanimous collective agreement between the beneficiaries, in addition to exchanges between them and the producer, may prevent the latter from knowing that he can extract extra prices which would cover the extra cost of producing more, or from achieving this extraction (the extra contribution that each beneficiary is ready to give usually does not by itself cover the cost); this is a cause of underproduction of the public good. But overproduction is also possible, because the producer may attribute some of the beneficiaries' global willingness to pay which he extracts to their willingness to pay for the last units.

In addition, exclusion from the benefit of the public good may not be possible. More generally, it has costs, difficulties or inconveniences which at some point deter the producer from excluding or the consumer from demanding access to the benefit of the good. These are of the following kinds: (1) Costs to the excluder such as doors, walls, fences, screens, guards, scrambling devices for herzian waves, hiding information, suing 'trespassers', etc. (2) Costs to users by inconvenience in consumption, use or benefit, as would for instance be the case of tolls on urban streets. (3) Costs to users by co-exclusion from other services, a case to be found notably when exclusion would be achieved by excluding from a complementary consumption, an important example being exclusion from a location (on private or public space) where the collective benefit can be received. (4) In the latter cases, the producer of the public good may not have the right to exclude from these other services (without problems 3 and 4 exclusion could always be made, for instance by banishment or even killing). (5) Logical impossibility such as when the desire for the consumption requires experiencing the consumption so that exclusion destroys demand, or the related situation described below in the very important case of 'collective gifts'.

If exclusion is not possible, or is too difficult or costly, production through individual initiative does not occur. Whatever the situation of excludability, the solution is a unanimous collective agreement between the beneficiaries and the producer(s), as to the good to produce and to each beneficiary's contribution to its financing. It is right and legitimate if the concerned persons use only legitimate rights to reach and achieve the agreement. But two things may prevent its spontaneous emergence. Firstly, if exclusion from the benefit of this public good is impossible or too difficult, a beneficiary is induced not to take part in the collective agreement, so as to benefit from the good produced thanks to others without contributing (he would be a 'free rider'); the result is inefficient, and if each beneficiary does that, relies on the others and is relatively small, no quantity of the good is produced at the limit. Secondly, whatever the situation as to exclusion, the transaction and information costs necessary to reach the agreement may prevent its achievement – all the more so when the number concerned is large. In these cases, there arises an ethical duty for a public service to achieve as far as possible what these free agreements would have done. This implies both the production of the public good and paying for it in the form of taxes.

This principle indicates both what quantity of the good must be produced and what the taxes must be.

This implicit exchange implies that the good must be produced up to the point where what the beneficiaries are ready to pay for an extra unit ceases to cover the cost of the latter (Dupuit's condition). It also implies that each beneficiary must not contribute to the payment more than the monetary equivalent of its value to him, whereas their total contribution must not fall short of total cost. This is ‘benefit taxation’, but for implementation this tax can either be levied specifically in relation to this advantage (and eventually be ear-marked for the provision of this good), or be aggregated to other notional taxes paid by the person for other specific reasons, i.e., into a more general tax scheme. Furthermore, this principle is not sufficient since it only gives an upper limit for each of the taxes and a lower limit for their sum – and it does not say how the) ‘surplus’ due to the provision of this public good is allocated. The implicit agreement principle however implies that the surplus is allocated between the putatively contracting parties, the beneficiaries and the producer(s) of the good (for its primary allocation, since general economic interdependence induces further effects). Not infrequently the sole information that taxes must not exceed willingnesses to pay and must not in total amount to less than cost defines them relatively precisely, in particular when beneficiaries have alternatives to the consumption of this public good. For the general case, the levels of these taxes, or the distribution of the surplus, are determined thanks to the bargaining theory which is relevant to the situation. This theory relies on the fact that the pre-agreement allocation and distribution is a legitimate starting point from the very definition of the general ethical theory. It deduces, in general, that the surplus must be allocated in relation to the agents’ legitimate bargaining power. The latter is itself determined by the fact that the solution must not be such that a group of persons can prefer to leave the general agreement and eventually build and finance a public good for themselves. The members of this group are influenced by the choices of the others (in opposition to the usual assumption in the concept of the) ‘core’), and in particular by what the others decide to produce of the public good if these ‘dissidents’ can still benefit from it. The result thus depends very much on this structure and on the possibility of ‘exclusion’ from the benefits of the good, but the general ethical theory leads to the adoption of a solution which assumes the possibility of exclusion (i.e., which disregards its difficulties) as a principle of what the taxes and the surplus allocation must be.

The practical application of the theory to determine the right amounts of public goods and the corresponding taxes requires the estimation of agents’ willingness to pay and potential bargaining power. A whole array of various methods is available, and more or less used, to gather information about these variables (Kolm 1972–74, 1973a, 1974b). For the monetary values, knowledge of technology and production functions provide the answer when agents are firms or when the public good benefits consumers through the technical production of some more final good. These values can also be inferred from the observation of market values of consumptions which are substitute or complementary to the public good (for example private material protection for public police protection, land values or rents for environmental quality differences, wage as value of time to estimate the worth of improved transportation infrastructure, etc.). Furthermore, in a poll asking people about the value of the public good to them (or of a variation of its quantity), they have no incentive to lie if they perceive that the public good produced and what they will pay are unrelated to their answer. If the beneficiaries are numerous, the relatively small ones know that their own answer will practically not influence the quantity chosen and, therefore, their payment through this way (except, perhaps, if it is considered as representative in a small sample). This allows us to gather easily the information necessary to produce the good, but not the one to set individual taxes (we would for instance gather the answers in an anonymous way). However, empirical experiences show that, de facto, people’s answers do not differ much according to whether their payment is tied or not to their answers (Bohm 1972). The political process is also often a notable source of information on the variables under consideration. And quasi-political processes can also be set in order to help solve the problem; for instance, the right quantity of public good and corresponding taxes, compared to nothing, gain unanimity in a referendum.

Several types of social situations and psychosocial phenomena help to secure the spontaneous collective realization of public goods or, at least, the relatively truthful revelation of preferences for them. Simultaneous contributions with each contributor watching the others contribute are possible in small groups, and, even in somewhat larger ones, the recurrence of similar situations allows retaliatory threats against free riding. But the most important are social norms and ethical behaviours (which are similarly necessary for private goods through the respect for property which establishes free exchange rather than widespread theft). These norms and behavioural patterns are acquired through socialization or, eventually, selection, and they tend to act better when the individual costs are either relatively small (voting, non-polluting) or dramatic (individual or collective danger). They include truth-telling, Kantian categorical imperative, altruism, communitarian feeling, internalized or approved good citizenship, imitation, the 'helping behaviour'studied by social psychology, gratitude and return gifts, general reciprocity, etc. Economists have also devised a series of 'revelation mechanisms' where a public centre receives information from the consumers of the public good and imposes on them incentives so as to induce them to 'reveal' the information about their tastes or needs which is necessary for the efficient production of the good; such devices have been proposed by Vickrey – in a different context – Kolm, Drèze and de la Vallée Poussin, Malinvaud, Groves, Ledyard, Clarke, Tideman and Tullock, Green and Laffont, and a number of others; although some of these mechanisms might eventually find some applicability, their general efficiency still requires that the agents have some motivation of the mentioned categories, beyond exclusive restricted self-interest.

The economics of public goods is as old as that for private goods (aside from literary remarks such as those of David Hume about public goods). In Paris in 1838, when Cournot was drawing a demand curve for a private good to theorize monopoly exploitation, Jules Dupuit drew one for a public good in order to choose the socially best quantity and financing of public works. Dupuit's 'demand curve' for a public good gives the number of users for each level of a putative toll (if they could be excluded) – i.e., the number of beneficiaries who value the good at more than this level. Equivalently, it is the distribution curve of the willingnesses to pay for this benefit classified in decreasing order. Dupuit's) 'surplus', the area under this curve, is thus the sum of all these willingnesses to pay. Dupuit's rule was to implement a variation in the specifications of the good as long as the surplus for this variation exceeded the cost of implementing it. Dupuit then set out to discuss how to discover willingnesses to pay, noting in particular its correlation with capacity to pay as revealed (in part) by the evident belonging of the user to a social class.

The history of ideas then leads us through the very perceptive Italian school of public finance (Mazzola, Pantaleoni, etc.) and to Wicksell's and Lindahl's 'just taxation' propositions, both of which based, on a clear conception of joint consumption and quantity optimality, illjustified fiscal schemes. (Wicksell's global public budget chosen at unanimity – 'quasi' for practicality – violates the right of a group of citizens to receive a public good and pay the corresponding taxes even if its effects displease other citizens, as long as it does not violate their legitimate rights. As for Lindahl, his pricing at marginal willingness to pay multiplied by quantity has no more justification than many others and its formal mimic of private good optimality formula is irrelevant. See these authors' articles in Musgrave and Peacock 1958.) Later, Bowen (1943) tried to achieve the optimum through a uniform given charge plus majority voting on the amount of the good, and Samuelson derived the quantity optimality condition from a welfarist, utilitarian, ethic with free lump-sum income redistribution.

The observation that the world exhibits few 'pure' public goods, whereas collective concern is widespread, led to the analysis of the various mixtures of privateness and publicness, and of their consequences for public finance and action. A 'good' may have to be defined by several parameters, some of which are privately divided whereas others are collective concerns. In fact, a uniform quality of a private good is technically a 'public good'. Increasing returns to scale in production is due to inputs which collectively serve several units of output, and this takes us to the classical public good case when this output is divided between several users (or even measured by the number of users served). A good can be privately divided between collectivities each of which consumes it collectively (this is usual for the intermediary private commodities entering into the production of several different public goods, and an example is the case of 'local public goods' the benefit of which is limited in space). Conversely, a good can be consumed collectively by groups each of which divides it between its members (an example is that of a given space which is fully occupied successively in time by different groups of persons). Different agents can benefit more or less from a public good according to some characteristic (location is an example). Or an extra consumption may neither crowd out another one as for private goods, nor leave other consumers indifferent as is the case with pure public goods, but just decrease some quality of the situation (as does for instance congestion and relative crowding). Or, also, uniform prices are collective concerns, with the corresponding optimality formula. (A systematic analysis of these pervasive intermediary situations with optimality conditions can be found in Kolm 1969a, b.)

Collective Gifts and Public Redistribution and Transfers

A liberty-based social ethic implies that legitimately held rights and property can be legitimately transferred by gift. The most common character of benevolence is that a single person is ready to pay something in order for someone in need to receive one dollar more, but far less than one dollar. However, it is commonly the case that several persons, often many, have a disposition of this type towards the needy individual. If the sum of what people are ready to pay for this specific person to receive one dollar more exceeds one dollar, a set of such transfers will be made by free agreement. The welfare of the person helped is a public good for the givers, and this situation is a special case of the public good questions. However, this category of public good is bound to be in a particularly unfavourable situation for spontaneous realization. Firstly, it is a case where exclusion is impossible for a purely logical reason since consuming the public good is knowing the needy person's situation, so that exclusion is hiding it or keeping another in ignorance of it, and then the potential giver is no longer ready to give (he does not give to know but to improve, if he does not know he does not give to improve, but if he knows he is not excluded). Secondly, the number and dispersion of co-givers and potential receivers creates information and transaction difficulties which impair direct agreements (this aspect, though, contrary to the 'free rider' aspect, could partially be met by private charitable institutions, which would however tend to be in undersupply because they cannot be profit-seeking). These transfers are thus an essential task of the public sector. As in the general public good case, each payer must be forced to pay the corresponding optimal taxes although he prefers the whole system of these redistributive transfers to its absence.

The determination of these transfers uses the whole gamut of means described in the general public good case. In addition, an important source of global information about them consists of analysing the reasons which motivate the co-givers, since these are often cultural and held in common. Beyond the basic relief of wretchedness, some of these reasons are more subtle, yet very important and widespread.

One of them is 'fundamental insurance'. Social insurance against basic life contingencies (disease, old age, cost of children, unemployment, etc.) presently often use as much or more money than public budgets not counting them. They are mutual insurance schemes which a priori can be private and frequently are. The basic reason why they are often public involves the idea of 'fundamental insurance'. This is a putative mutual insurance against causes of hardship which happened before an effective insurance could be taken out by the person, such as proneness to disease, genetic disposition, or poor education or motivation resulting in low income in the labour market, etc. The idea is that people who happen to have such 'bad luck' or 'misfortunes' must be helped by the more ‘lucky’ or ‘fortunate’ ones – in addition to the insurance consequences of occurrences which happen later. Since insurance against these previous events (allocation of natural abilities, genes or education) cannot be effectively taken out by the existing adults, this market is missing and the public sector must supplement it.

A number of insurance markets for future contingencies are also missing for the more standard difficulties of envisioning and defining the risk (this often happens, for instance, for economic downturns and unemployment or for major natural catastrophies). This provides a reason for a second category of ‘implicit insurances’ achieved by public transfers to people incurring the specific misfortune.

More generally, transfers to people in need, for reasons of helping, solidarity or fairness, also imply some equalization of situations, and egalitarian feelings are very common in opinions about public policy actions. These must however be carefully defined, scrutinized and sorted out. Freedom is equality of basic rights. But it so happens that transfers from rich to poor, when no mention is made of the cause of the disparity, and when no disincentive indirect effect makes everybody finally worse off, has an appeal wider than it may seem, as various people approve of diverse redistribution schemes which are equivalent to it although that is not apparent (Kolm 1966, in Margolis and Guitton 1968–9). But this has an applicability limited to the distribution of income without previous legitimacy either from pure chance or from some natural resources. On the other hand, feelings of envy or jealousy are not commonly regarded as ethically defensible reasons for transfers (but equality of opportunities, a principle of wide acceptance, leads to general no-envy and no-jealousy since each person had the opportunity to choose what each other has chosen and he has preferred his own choice).

Finally, a recurrent and practical issue about public transfers and assistance is whether they should be in cash or in kind. The decision here rests with the givers who initially own the funds. The arguments are for them to consider, and if cash has the advantage of leaving the receiver free to choose what he prefers, they must be convinced of it.

Externalities

If a person’s action or situation importantly concerns another person without there being an exchange (or a chain of exchanges) between them so as to adjust their interlocking desires, there results (inefficient) ‘externality’ and a duty for the public sector to implement what this exchange would have been (from legitimate rights), were it not for the information or transaction difficulties or costs, or uncertainties about rights, which prevent its spontaneous existence. The corresponding compulsory money transfers can be either way between the creator and the receiver of the externality: for an external diseconomy it is a tax on the creator and a compensation to the receiver, or a subsidy to the creator who abstains somewhat, paid by the receiver, and for an external economy it is a payment to the creator from the receiver, or a tax on the creator, who does not create as much as he can be required to, and a payment to the receiver. (The result is Pareto-optimal but, in opposition to the Old Public Economics, ‘welfarist’, utilitarian or ‘efficiency’ view, firstly just to impose the level of the action will not do – there must be compensation; secondly, the tax on or subsidy to the producer of the externality alone will not do – it must be paid to or from the victim or beneficiary, and this even is the aim of the operation, Pareto-optimality being a secondary consequence of it; and thirdly, the reference situation defining the base of the tax or subsidy is a definite right and not any indifferently arbitrary point.)

Quite often, an external effect concerns several, or many, persons together, thus mixing externality and public good problems. Conversely, joint production of an effect exists when one cannot identify the specific producer of it (as in car pollution). These collective aspects bring in ‘free rider’ and transaction costs problems which often prevent direct agreement, thus creating the situation of externality rather than exchange. Such phenomena, along with the question of the allocation of new environmental rights emerging from scarcities created by economic growth, are the core of the problems and theories of the economics of public environment, an outgrowth of Public Economics, buoyant since the early 1970s.

Extensive studies have been devoted to a special but important case, which is also an essential social phenomenon. This is the situation where a creator of an externality puts himself by the same act in the situation of being receiver of an externality of the same kind. These are called *congestion externalities* (since this happens particularly in situations of congestion). The problem is that if someone is compensated for the external diseconomies he incurs (or pays for the external economies he creates), this may erase his having to pay for the diseconomies he produces (or what he receives for his external economies), so that the relevant transfers are not possible or are only imperfect, and in particular they cannot induce efficiency. The solution is to devote the product of the tax to improve global quality (for instance more room in the case of congestion) which is a public good the correct amount of which does not depend upon the decision of any single individual if he is relatively small enough (the financial result depends upon the basic structure of 'qualitative return to scale': Kolm 1969b, 1974a).

Rectifications

A consequence of the public sector's duty to protect rights is its duty to rectify past violations of legitimate rights. Such judgements are commonly considered by the courts, but one may have to rectify, at least in a global way, the effects of more ancient and deeper violations (given a statute of limitations provision). The resulting actions are compulsory transfers and, eventually, some specific measures like education.

The Theory of the Public Debt

The public debt is the means of making *retro-payments*, that is, payments when the payer exists later in time than the receiver: the receiver receives from the public sector which pays out of public borrowing, and the payer pays by later taxes used to redeem the public debt and meet its interest. Time introduces two kinds of constraints on the relations between persons at different dates, hence two causes of 'market failures'. Nothing can be received before it is handed out, and the individuals involved in the exchange cannot make an explicit mutual agreement; the only possible relation is forward transfer (through asset accumulation) decided by the transferer, that is, a forward gift. From its general principle, the public sector must abolish these constraints on free relations if it can. Indeed, it can make retro-payments through public debt, and it can attempt to evaluate the desires of future generations. This opens up the possibility of publicly implemented retro-buying intertemporal exchange and retro-gift. Public debt enables one to buy the earlier production of a durable commodity which yields benefits he desires (to him or to others), and it enables one to make at some date a gift which helps satisfy previously existing needs of other people. For instance one can in this manner buy durable improvements (or protection) of the environment, and also give to victims of a previous bad specific or general situation. Of course, even if this buyer or giver desires this, his payment must be compulsory – i.e., it is a tax, since when it occurs the first payment has already been made. This reason for public intervention and for public debt often intervenes with the other reasons for public intervention: the benefit taxation to finance a durable public good implies public borrowing later redeemed thanks to contributions of beneficiaries which actualize their implicit basic agreement with the beneficiaries at other dates; a global reflation through deficit finance, when it works, is paid for by the future beneficiaries of higher employment, demand and capital formation, through taxes which will redeem the debt; there are also collective retro-gifts, inter-temporal implicit – and in particular fundamental – insurances, and intertemporal allocation of the values of natural resources.

Macroeconomic Policy

Another category of market failure has such pervasive effects and implications that a collective rights-preserving unanimous agreement as to corrective action would involve a prohibitively large number of persons, so that the public sector has to intervene. Such phenomena are those which result in mass unemployment and inflation, with possible effects as to large indebtedness to foreigners or slow productivity growth. (Inflation and insufficient global demand are public 'bads'. In the case of inflation, the primary reasons are psychosocial effects which disrupt social predictability, general confidence, and the implicit social contract, due to uncertainty and volatility of relative prices; see Kolm 1984a.) The 'market failures' which cause them involve a whole array of informational limitations, the non-existence of pure market equilibria (at least of those in which no one starves), the lack of a spontaneous tendency to equilibrium of a multimarket system, the income effects of wage decreases which diminish income and employment, the general absence of futures markets which would have guided investment decisions, downward price inflexibilities due to status and norm effects of wages and prices and to average asymmetries in the optimal allocation of selling and buying efforts, monopolistic situations and collusions (in labour markets for instance), etc. Apart from some direct microeconomic income-price or employment or other policies, the corresponding corrective tools are those of macroeconomic policy; that is, arrangements between public expenditures, taxes, public debt, money supply, foreign money, through decisions about quantities or rates (rates of interest, rates of foreign exchange, and taxes).

The Right Taxation

What the public sector must do is the whole of all the specific actions described above. In particular, ethical consistency in general requires taxes to be justified by expenditures. Institutionally and practically, however, taxes can be pooled and levied more or less as a single tax or as a relatively small number of taxes. The corresponding economy of administration must be balanced against the advantages of decentralized and specialized public services financed by the logically ear-marked right taxes, regarding the budget allocated to this service, the closeness of the public management to the users and its better knowledge and awareness of their needs; the better understanding by the taxpayers of the use of their money and of the benefit they derive from it and, therefore, their greater possibility of checking the use of public funds and of controlling the public service by political means.

However, the above reasons for specific constitutive taxes imply some general structures of the global taxation faced by the individuals. A large number of specific taxes (in the sense of justified tax elements), taken as function of income, have the form of an increasing tax (often a relatively proportional one) above some exemption level. This tends to be the case for public goods which yield benefits according to some ownership or activity, for redistributive transfers due to collective gifts and in particular to fundamental insurance, for rectification compensations. The summing up of all these taxes therefore yields a progressive tax schedule, and even approximately a succession of tax 'brackets' with flat tax rates increasing from one to the next. Furthermore, collective gifts, fundamental insurance and rectification for past infringements of rights, globally conduce to a negative taxation scheme at low income levels (i.e., people whose incomes fall below a certain level receive an assistance which increases with this gap).

Public Prices, Taxes and Production

Taxes must be based on observable facts, which practically will often mean quantities or values of commodities of many possible kinds, including incomes – whether these are the right base or only proxies adopted to confront informational or management difficulties. To choose these taxes is thus equivalent to choosing the corresponding prices. Now, to choose prices is also a problem the government faces in the case of public firms or regulated industries. In addition, a standard reason why firms are public or controlled is the existence of strong increasing returns to scale, so that competitive prices would bring them into deficit, and a private monopoly with limited discriminatory power would produce inefficiencies; yet these are financially autonomous firms which must not have a permanent deficit. Therefore, the problems of choosing the socially best taxes in order to provide a public income, and that of choosing the socially best prices respecting a budget constraint, are de facto identical. This problem was extended in a general 'theory of value constraints', which are constraints on elements of budget (global or partial or ratios) and on prices. In addition, the choice of a non-linear public tariff applicable to several users is isomorphic to that of an income tax structure. Determinations of optimality conditions and structures in all these cases involved the successive works of Colson (1901–5), Divisia (1927), Ramsey (1927), Boiteux (1956), Kolm (1969), Mirrlees (1971), Diamond (1975), Atkinson and Stiglitz (1980), Guesnerie (1980) and others.

Situation of Public Economics

Public Economics is probably for many reasons the most paradoxical field of economics. It specializes in what has always been the central problem of the discipline – whether it was attacked frontally or treated in a devious, consequential and somewhat hypocritical way: what must be market, what public, and how? Yet it often works at the borderline of the profession, flirting more and more with moral philosophy, political science, organizational studies (public choice and the economics of institutions), with breakthroughs to come with psychology ('multidimensional man') and sociology. It specializes in the public sector, yet it must for this reason involve the finest analysis of the market and of its problems. It has a well-defined, specific style and flavour, yet it is hard to circumscribe rationally: for instance does it contain Macroeconomics – which comes more and more to the core of Public Economics as it relies more and more on microeconomic analyses of market failures? What is its relation to International Economics which studies markets but owes its existence to that of States? It probably contains Public Finance in scope, yet the latter's tradition has a quite different, less analytical, style; and we could go on with Public Choice, Social Choice, the theory of constitutions, 'Political Economics', the theory of bureaucracy, Welfare Economics whether theoretical or applied, the essential benefit-cost analysis, the economics of socialisms, comparative economic systems, etc. Public Economics is founded on positive views of the public economy and of the market, yet its final aim is almost always normative. It is one of the oldest fields in economics, yet it also is one of its newborns.

These distinctions, however, are interesting but not essential. The thing is to take an important problem and to find powerful tools to crack it. Public Economics' central question – to know what the public sector must do, when the markets work better, or if a hopeful 'third sector' and reciprocity relationships would be a still better possible solution – can hardly see its import challenged. As for the tools, it started with sharp economic analysis, but saw its utility impaired by ethical naiveté it is now incurring a new boom in useful results thanks to an equivalent input of social ethics; as for the future, the next threshold will probably be to work out the consequences of a much richer and more truthful account of man's motives and capacities, along with an understanding and a consideration of institutions, both meaning psychology since it is how institutions enter into an individualistic social theory: through people's heads and hearts.

See Also

- ▶ Collective Action
- ▶ Constitutional Economics
- ▶ Public Finance

Bibliography

Andreé, C., and R. Delorme. 1983. *L'état et l'économie*. Paris: Seuil.

Arrow, K.J., and T. Scitovsky. 1969. *Readings in welfare economics*. London: George Allen & Unwin.

Atkinson, A., and N. Stern. 1974. Pigou, taxation and public goods. *Review of Economic Studies* 41: 119–128.

Atkinson, A., and J. Stiglitz. 1980. *Lectures on public economics*. New York: McGraw-Hill.

Bator, F. 1958. The anatomy of market failure. *Quarterly Journal of Economics* 72: 331–379.

Baumol, W. 1952. *Welfare economics and the theory of the state*. London: Longmans Green.

Baumol, W.J., and W.E. Oates. 1975. *Theory of environmental policy, externalities, public outlays and the quality of life*. Englewood Cliffs: Prentice-Hall.

Bénard, J. 1985. *Economie publique*. Paris: Economica.

Bobe, B., and P. Llau. 1978. *Fiscalité et choix économiques*. Paris: Calmann-Lévy.

Bohm, P. 1972. Estimating demand for public goods: An experiment. *European Economic Review* 3: 111–130.

Boiteux, M. 1956. Sur la gestion des monopoles publics astreints á l'équilibre budgétaire. *Econometrica* 24: 22–40.

Bowen, H. 1943. The interpretation of voting in the allocation of economic resources. *Quarterly Journal of Economics* 58: 27–48.

Buchanan, J.M. 1965. An economic theory of clubs. *Economica* 32: 1–14.

Buchanan, J. 1968. *The demand and supply of public goods*. Chicago: Rand McNally.

Buchanan, J., and G. Tullock. 1962. *The calculus of consent*. Ann Arbor: University of Michigan Press.

Clarke, E.H. 1971. Multipart pricing of public goods. *Public Choice* 11: 17–33.

Colson, C. 1901–5. *Cours d'économie politique professé à l'Ecole Nationale des Ponts et Chaussées*. Vols. 1,3,5,6. Paris: Gautheir-Villars.

Diamond, P. 1975. A many-person Ramsey tax rule. *Journal of Public Economics* 4(4): 335–342.

Diamond, P., and J. Mirrlees. 1971. Optimal taxation and public production: Pts. I and II. *American Economic Review* 61: 8–27; 261–278.

Divisia, F. 1927. *Economique rationnelle*. Paris: G. Doin.

Divisia, F. 1937. *Cours d économie politique et sociale*. Paris: Ecole Polytechnique.

Drèze, J., and Poussin De la Vallée. 1971. A tâtonnement process for public goods. *Review of Economic Studies* 38: 133–150.

Dupuit, J. 1844. De la mesure de l'utilité des travaux publics. *Annales des Ponts et Chaussées* 8(116): 332–375 (2nd series).

Dupuit, A.J. 1849. Sur les péages et le prix des transports. *Annales des Ponts et Chaussées.*

Ekelund, R., and R. Hébert. 1973. Public economics at the Ecole des Ponts et Chaussées. *Journal of Public Economics* 2(3): 241–256.

Feldstein, M., ed. 1976. *The economics of public services*. Proceedings of Turin I.E.A. Conference. London: Macmillan.

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

Frey, B. 1985. *Economie politique moderne*. Paris: PUF.

Green, H. 1962. The social optimum in the presence of monopoly and taxation. *Review of Economic Studies* 29: 66–78.

Green, J., and J.J. Laffont. 1976. Révélation des préférences pour les biens publics. *Cahiers du Séminaire d'Econométrie*. Paris: CNRS.

Green, J., and J.-J. Laffont. 1979. *Incentives and public decision making*. Amsterdam: North-Holland.

Groves, T. 1973. Incentives in teams. *Econometrica* 41.

Groves, T., and J. Ledyard. 1977. Optimal allocation of public goods: A solution to the 'free rider' problem. *Econometrica* 4: 783–809.

Groves, T., and M. Loeb. 1975. Incentives and public inputs. *Journal of Public Economics* 4: 211–226.

Guesnerie, R. 1980. *Modèles de l'économie publique*, Cahiers du séminaire d'économétrie. Paris: CNRS.

Guillaume, H. 1973. *Prix fictifs et calcul économique public*. Paris: Monographie du séminaire d'économétrie.

Hahn, F. 1973. On optimum taxation. *Journal of Economic Theory* 6: 96–106.

Haveman, R. 1976. *The economics of the public sector*, 2nd ed. New York: Wiley.

Haveman, R., and J. Margolis. 1977. *Public expenditures and policy analysis*. Chicago: Rand McNally.

Head, J.G., and C.S. Shoup. 1969. Public goods, private goods and ambiguous goods. *Economic Journal* 79: 567–572.

Hotelling, H. 1938. The general welfare in relation to problems of taxation and of railway and utility rates. *Econometrica* 6(3): 242–269.

Hume, D. 1739. *A treatise on human nature*. Oxford: Oxford University Press, 1895.

Jessua, C. 1968. *Coûts sociaux et coûts privés*. Paris: PUF.

Johansen, L. 1965. *Public economics*. Amsterdam: North-Holland.

Johansen, L. 1977. The theory of public goods: Misplaced emphasis? *Journal of Public Economics* 7: 147–152.

Kolm, S.-C. 1964. *Les foundements de l'économie publique, introduction à la théorie du rôle économique de l'état*. Paris: IFP.

Kolm, S.-C. 1969a. *Cours d'économie publique*, L'état et le système des prix, vol. I. Paris: Dunod.

Kolm, S.-C. 1969b. *Cours d'économie publique*, Le service des masses, vol. II. Paris: Dunod.

Kolm, S.-C. 1971. *Justice et équité*. Paris: CNRS.

Kolm, S.-C. 1972–4. Ascertaining environmental costs and benefits. In *Environmental damage costs*. Paris: OECD.

Kolm, S.-C. 1973a. *Estimation des coûts et valeurs des biens publics*. Paris: Rapport au Commissariat général du Plan.

Kolm, S.-C. 1973b. Super-equité. *Kyklos* 26(4): 841–843.

Kolm, S.-C. 1974a. Le rendement qualitatif et le financement optimal des politiques d'environnement. *Econometrica* 56(5): 1056–1062.

Kolm, S.-C. 1974b. Connaissance des coûts et valeurs d'environnement. In *Les coû ts des dommages causés à l'environnement*, 22–25. Paris: OECD.

Kolm, S.-C. 1984a. *Sortir de la crise*. Collection Pluriel. Paris: Hachette.

Kolm, S.-C. 1984b. *Le libéralisme moderne*. Paris: Presses Universitaires de France.

Kolm, S.-C. 1984c. *La bonne économie, la réciprocité générale*. Paris: Presses Universitaires de France.

Kolm, S.-C. 1985. *Le contrat social liberal*. Paris: Presses Universitaires de France.

Laffont, J.-J. 1982. *Fondements de l'économie publique*. Paris: Economica.

Lesourne, J. 1964. *Le calcul économique*. Paris: Dunod.

Malinvaud, E. 1972. Prices for individual consumption, quantity indicators for collective consumption. *Review of Economic Studies* 39: 385–406.

Marchand, M., P. Pestiau, and H. Tulkens (eds.). 1984. *The performance of public enterprises: Concepts and measurement*. Amsterdam: North-Holland.

Margolis, J., and H. Guitton, eds. 1968–9. Proceedings of the 1966 Biarritz IEA-CNRS conference. *Economie Publique*. Paris: CNRS/*Public Economics*; London: Macmillan.

Matthews, R., and G. Stafford (eds.). 1982. *The grants economy and collective consumption*. London: Macmillan.

Committee, Meade. 1978. *The structure and reform of direct taxation*. London: Allen & Unwin.

Milleron, J.-C. 1972. Theory of value with public goods: A survey article. *Journal of Economic Theory* 5: 419–477.

Mirrlees, J. 1971. An exploration in the theory of optimum income taxation. *Review of Economic Studies* 38: 175–208.

Musgrave, R. 1959. *The theory of public finance*. New York: McGraw-Hill.

Musgrave, R. 1969. *Fiscal systems*. New Haven: Yale University Press.

Musgrave, R., and A. Peacock (eds.). 1958. *Classics in the theory of public finance*. London: Macmillan.

Nozick, R. 1976. *Anarchy, state and utopia*. New York: Basic Books.

Oates, W. 1972. *Fiscal federalism*. New York: Harcourt, Brace, Jovanovich.

Olson, M. 1965. *The logic of collective action*. Cambridge, MA: Harvard University Press.

Peacock, A., and F. Forte. 1981. *The political economy of taxation*. Oxford: Blackwell.

Pigou, A. 1928. *A study in public finance*, 3rd ed. London: Macmillan, 1947.

Ramsey, F. 1927. A contribution to the theory of taxation. *Economic Journal* 37: 47–61.

Rees, R. 1968. Second-best rules for public enterprise pricing. *Economica* 35: 260–273.

Roy, R. 1942. *De l'utilité – Contribution à la théorie des choix*. Paris: Hermann.

Roy, R. 1947. La distribution du revenu entre les divers biens. *Econometrica* 15: 202–225.

Samuelson, P. 1954. The pure theory of public expenditure. *Review of Economics and Statistics* 36: 387–389.

Sandmo, A. 1976. Optimal taxation: An introduction to the literature. *Journal of Public Economics* 6: 37–54.

Schotter, A. 1985. *Free market economics: A critical appraisal*. New York: St Martin's Press.

Terny, G. 1971. *Economie des services collectifs et de la dépense publique*. Paris: Dunod.

Tideman, M., and G. Tullock. 1976. A new and superior process for making social choices. *Journal of Political Economy* 84: 1145–1159.

Tiebout, C. 1956. A pure theory of local expenditures. *Journal of Political Economy* 64: 416–424.

Tinbergen, J. 1956. *Economic policy: Principles and design*. Amsterdam: North-Holland.

Tinbergen, J. 1958. *On the theory of economic policy*. Amsterdam: North-Holland.

Tulkens, H. 1978. Dynamic processes for public goods: An institution-oriented survey. *Journal of Public Economics* 9: 163–201.

Turvey, R. 1971. *Economic analysis and public enterprises*. London: Allen & Unwin.

Vessilier, E. 1977. *Economie publique*. Paris: Masson.

Vickrey, W. 1961. Counterspeculation, auctions and competitive sealed tenders. *Journal of Finance* 16: 8–37.

Wolfelsperger, A. 1975. *Les biens collectifs*. Paris: PUF.

Public Finance

Richard A. Musgrave

Abstract

Public finance may well be the oldest branch of economics. It concerns not only the effects of fiscal operations on the market but also principles of public sector economics, which address a distinct set of issues and are linked closely to the perspectives of political and social science. This article covers (a) expenditure on public goods and transfers, including its macroeconomic effects as analysed by the classical and the Keynesian schools, and (b) taxation, including issues of tax equity and tax efficiency, definitions of income, consumption and expenditure, and tax shifting and incidence.

Keywords

Ability to pay tax; Ad valorem taxation; Aggregate demand; Bentham, J.; Bequests; Budgetary policy; Capital budgeting; Capital gains taxation; Capital measurement; Consumption base; Consumption taxation; Cost–benefit analysis; Cournot, A.; Deadweight loss; Depreciation; Edgeworth, F. Y.; Elasticity of substitution; Entitlement; Equal absolute sacrifice theory of tax equity; Equal marginal sacrifice theory of tax equity; Equal proportional sacrifice theory of tax equity; Equality; Excess burden of taxation; Exchange; Expenditure base; Expenditure tax; Externalities; Factor mobility; Factor pricing; Fairness; Fiscal location; Free rider problem; Hobbes, T.; Horizontal and vertical equity; Human capital; Hume, D.; Income base; Income effect of taxation; Income subsidies; Inflation; Intergenerational equity; International capital flows; Interpersonal utility comparisons; Interpersonal utility interdependence; Justice; Keynesianism; Land tax; Leisure; Lindahl, E. R.; Local public finance; Marginal rates of substitution; Marginal utility of income schedules; Market failure; Merit goods; Mill, J. S.; Musgrave, R. A.; Neoclassical growth models; Neoclassical synthesis; Net income; Opportunity cost; Optimal commodity taxation; Optimal taxation; Paternalism; Payroll tax; Physiocracy; Pigou, A. C.; Preference revelation; Primary distribution; Progressive and regressive taxation; Propensity to save; Property tax; Public choice; Public expenditure; Public finance; Public goods; Public policy failure; Quasi-rent; Rawls, J.; Rent; Retail sales tax; Ricardo, D.; Samuelson, P. A.; Scandinavian model; Shadow pricing; Smith, A.; Social discount rate; Social welfare function; Subsistence hypothesis; Substitution effect of taxation; Tax avoidance; Tax base definition; Tax burden; Tax efficiency; Tax equity; Tax expenditures; Tax incidence; Tax prices; Tax shifting; Taxation of corporate profits; Taxation of income; Tiebout hypothesis; Transfers; Utilitarianism; Value-added tax; Voting rules; Wagner’s law; Wicksell, K.

JEL Classifications

H0

Concern with public finances may well be the oldest branch of economics. Statesmen have needed advice, and writing on fiscal affairs dates back to antiquity. It concerned the scholastics of the 16th century and occupied the mercantilists of the 17th. Systematic study of the public household by the Cameralists followed, and the ‘L’impôt unique’ was a central part of Physiocrat doctrine. In England, the writings of Petty, Locke, and Hume preceded Book V of Adam Smith’s *Wealth of Nations,* the first ‘modern’ statement of the field. Thereafter, fiscal analysis followed (and in some cases led) the advances of economic science. Ricardo, Mill, the marginalists, Marshall, Pareto, and Pigou all left their stamp on the economics of public finance, not to mention the impact of Keynes and the emergence of stabilization as a goal of budget policy. But fiscal economics also added to the general body of economic analysis. Its concern is not limited to the effects of fiscal operations on the market, and market responses thereto. There remains the more basic question of why a public sector is needed and what rules should be applied to its conduct. Principles of public sector economics are required to provide the answer. These principles, to be sure, are coordinated with those of the market by the broader frame of economic welfare, but they address a distinct set of issues and, by their very nature, are linked more closely to the perspectives of political and social science.

Public Expenditures

We begin with the expenditure side of the picture, and public goods in particular. Thereafter, transfers are examined.

Public Goods

Here the basic issue is why certain goods and services have to be provided for through the budget, i.e., paid for by taxes and made available free of direct charge. Such goods and services may be produced under public or private management, e.g., government may install traffic lights itself or engage a private firm. Which is done does not matter here. The crucial point is that traffic signals are provided free of direct charge to the individual consumer when passing the intersection. Given a general presumption that consumer preferences should be met by the *quid pro quo* of the market, why should budgetary provision be chosen in the case of public goods?

A modern answer was anticipated by Hume’s early insight. Two neighbours, so he argued, might agree to drain a meadow, but no such agreement can be reached by a thousand persons, as each will try to place the burden on others (Hume 1739, p. 539). Adam Smith also examined why certain services must be provided by the Prince (Smith 1776, Book V). These included the upkeep of the court, defence, police, and minimal education for the poor. More generally, the Prince was to provide for 'those institutions and public works which, though they may be highly advantageous to a great society, are, however, of such a nature that the profit could never repay the expense to any individual' (Smith 1776, vol. 2, p. 539). The issue was joined but Hume's earlier insight had been lost. The reason why private provision will not work remained unanswered. John Stuart Mill subsequently came a bit closer. He noted that individual preferences, under certain conditions, cannot be met without common concert and legal sanction. The lighthouse illustration appeared (Mill 1848, p. 976), and the difficulty of collecting tolls for the use of its services was stressed. Market failure due to inapplicability of exclusion was thus noted, but not as yet the more basic proposition that exclusion would be inefficient (absent crowding) even if it could be applied.

The discussion took a new turn in the 1880s, when utility analysis grounded value theory on the demand side. Public no less than private services are provided to meet the preferences of individual consumers, so it was argued, and this provision through the budget may be viewed in analogy to the exchange mechanism of the market. Taxes may be viewed as price payments, offered by the consumer. Thus a voluntary exchange model of the tax-expenditure process emerged (Sax 1887; Mazzola 1890). This vision of the fiscal process was rejected at once by Wicksell (1896). While public provision should be in line with individual preferences, it could not be implemented by voluntary exchange. As Hume had recognized a century and a half before, exchange will not work in the large number case. The level of public services available to A will not be affected significantly by his own contribution. Hence A will not reveal his preference for public services. Unlike the case of private goods, where the individual must bid to obtain his share in the auctioning process of the market, the consumer will now act as a free rider. A political process of budget determination by voting, combined with a legal sanction of its enforcement, is needed so that preferences are revealed. Thus the basis was laid for the modern discussion of public choice and the voting rules by which an efficient solution may be approximated. Though the simplistic hypothesis of voluntary exchange was rejected, the exchange formulation as developed by Lindahl (1919) nevertheless had an important role to play. With any given supply of the public service available to both A and B, A's demand curve may be viewed by B as a supply curve.

Equilibrium is then reached where the vertically aggregated demand curves of A and B intersect the supply schedule for the product, i.e., where the tax prices paid by the two consumers add up to the social cost of the service. Lindahl's formulation, with its vertical addition of demands, thus anticipated a significant feature of Samuelson's later formulation.

The next phase of the development of public goods theory emerged with Pigou's analysis of externalities (1920). External costs and benefits remain unaccounted for by the market and thus call for correction by public policy. Pigou did not develop this theme in his treatise on public finance (1928), where the proper sphere of public services is defined only in general terms, calling for extension to the point where marginal social costs and benefits will be equal. However, the concept of external benefits might be extended readily to that of public goods, where externalities appear not as a by-product of internal gains but all benefits are external. Introduction of the Scandinavian model into the English-language literature followed (Musgrave 1939), but the crucial implication of public goods for efficient resource use was not drawn clearly until Samuelson's statement (1954). Whereas efficient use of resources in the provision of private goods calls for an equating of the marginal rates of substitution in production and consumption (with the latter equal for all consumers), that of public goods calls for equality of the marginal rate of substitution in production with the *sum* of the marginal rates of substitution (differing among consumers) in consumption. Lindahl's earlier exchange solution, arrived at by the vertical addition of demand curves, was compatible with this outcome; but, as shown in Samuelson's model, an efficient solution did not require the Lindahl-type determination of tax shares.

The basic difference in the two formulations should be noted. Lindahl, following Wicksell, began with a distribution of money income and then proceeded to assign efficient tax prices. This is accomplished by charging each consumer in line with his marginal evaluation and setting total supply so as to equate the sum of these charges with cost. Moreover, by postulating a just state of distribution to prevail to begin with, a requirement also advanced by Wicksell, the resulting burden distribution of tax shares would also be just. Samuelson, like Wicksell, rejected Lindahl's 'pseudo demand curves' as unrealistic. But where Wicksell proceeded to examine the process of preference revelation, Samuelson provided a more general definition of the efficient solution. Preference revelation is disregarded as the model visualizes an omniscient referee to whom preferences are known. He then establishes a utility frontier, showing various mixes of public and private goods, as well as private good distributions among individual consumers. The optimum optimorum (bliss point) on the utility frontier is then chosen on the basis of a social welfare function. The Lindahl solution becomes a special case only; but given the assumption of known preferences, there is no particular reason why it should be selected.

This more general formulation resolves the allocation and distribution aspects of the problem simultaneously, and deals with distribution in its basic welfare rather than income terms. The Lindahl model by comparison separates allocation and distribution issues. Since welfare is a function of real, not money income, it is open to the objection that the just distribution of money income cannot be determined without also setting tax prices, thus suggesting circular reasoning. However, this critique may be met by adding the determination of the voting rule (designed so as to best secure preference revelation) as a further equation to the model. Moreover, the Lindahl formulation provides a closer linkage to the real world. While tax prices cannot be seen as voluntary offers, there exists no omniscient referee to whom preferences are known. Preferences, as Wicksell noted, must be revealed through a voting process and this presumes a distribution of money income to begin with. The Lindahl price thus remains as a benchmark against which the quality of the voting process can be measured. Separation of the allocation and distribution phase of budget policy was extended subsequently to include the stabilization function as a third branch (Musgrave 1959).

Subsequent developments in the theory of public goods recognized the fact that particular goods and services may not meet the polar conditions of purely private or public goods, but fall in between. A's consumption of X may provide benefits internal to A and hence be undertaken by him. But it may also generate external benefits or costs for B, C, and D. Thus, partial provision through the budget, that is, a subsidy solution, may be called for. Or, the number of consumers may be sufficiently small to permit a bargaining solution without voting; yet there is no assurance that an efficient outcome will emerge. Moreover, it may be possible to satisfy certain needs via the provision of public or of private goods, thus permitting a choice between the two modes.

A special problem arises also from the fact that the benefits of public services may be subject to spatial limitation. The resulting distinction between 'local' and 'national' public goods provided the basis for a theory of fiscal location (*see* "▶ Local Public Finance") according to which the provision of public services should be arranged so that each jurisdiction will provide and pay for the public services the benefits of which accrue within its borders. Moreover, the spatial limitation of benefits led to the proposal (Tiebout 1956) that preference revelation may occur through 'voting by feet'. Individuals with similar public goods preferences would find it advantageous to congregate. At the same time, individuals with lower incomes will find it advantageous to congregate with higher income individuals, so as to generate an unstable distribution across jurisdictions.

The general theory of public goods has thus been extended and qualified to deal with particular situations. It should be noted, however, that all these variants assume public goods to be provided in line with the preferences of individual consumers. The theory of public goods is thus similar in its psychological underpinnings to that of private goods. The concept of communal needs or goods which the community considers meritorious offers an alternative perspective not included in the mainstream view of expenditure theory (*see* "▶ Merit Goods").

Cost-Benefit Analysis

Moving from general theory to practical application, the development of cost–benefit analysis has attempted to design an operational framework by which the appropriateness of particular expenditure projects may be evaluated and ranked. In the process, the present value of the expected benefit stream is balanced against that of its opportunity cost. For this purpose, a social discount rate has to be determined, a rate which may or may not be taken to equal that of the market. Moreover, opportunity cost is found to depend upon whether the resource withdrawal is from consumption or from capital formation. Since the income tax enters as a wedge between gross and net return, capital formation in the private sector falls short of the optimal level so that resource withdrawal from private investment carries a higher social cost than does resource withdrawal from consumption. Shadow prices are applied to measure the social cost of labour and other inputs, thus correcting for further distortions in market prices. Finally, distributional weights, based on a social welfare function, may be applied to the resulting costs and benefits. Thus, a framework is provided by which the value of alternative expenditure projects may be assessed and ranked. However, the analysis assumes that the value of the benefit stream can be determined. This involves no difficulty where the output of the public project is sold at the market, but approximations have to be used where the services are in the nature of public goods. Thus, the value of a park may be measured by the opportunity cost reflected in the visitor's travel time or by similar proxies.

Transfers

While economic analysis has focused on the provision of public goods and services, transfers have come to claim an increasing share of total spending. Aimed at correcting the distribution of income, they may be viewed as negative taxes. While resource use for the provision of public goods may be fitted into the Paretian mould of allocation efficiency, transfers pose a more difficult problem. To be sure, a theory of giving, or Pareto optimal redistribution, may be developed in the context of interpersonal utility interdependence. If the donor's satisfaction is derived from the pleasure of individual giving, giving remains a private good. But if it is derived from the welfare of others, giving assumes a social-good quality and calls for budgetary provision. Yet the outcome reflects the initial distribution of income and thus does not resolve the more basic problem of primary distribution. This transcends considerations of Pareto efficiency, and broader grounding in a theory of justice is required, be it a Lockean rule of entitlement, a utilitarian concept of maximum welfare, or a Rawlsian sense of justice as fairness. But notwithstanding this inherent link to a theory of justice, economic analysis retains a decisive role. The size of the pie is linked to its distribution, and redistribution involves an efficiency cost. The one, therefore, cannot be determined without the other.

Turning to the form in which transfers should be given, economists have traditionally argued in favour of a general income subsidy, rather than selective subsidies designed to support particular uses of income. The general subsidy will be more valuable to the recipient as it does not interfere with his choice among income uses. But various exceptions may be noted. Transferors, consenting to a transfer, may do so on condition that the income is put to specified uses. Giving may take a paternalistic form. Moreover, distributive justice may be viewed in categorical terms, applying different standards of equality to basic items than to other income uses (*see* "▶ Merit Goods"). Beyond this, the logic of optimal commodity taxation, calling for differential rates of tax on various goods and services, may also be shown to call for differential subsidy rates to be applied to various commodities.

Macroeconomic Aspects

The preceding discussion has dealt with the role of public expenditures in providing for public goods and for adjustments in distribution. In the process, the fiscal system may add to capital formation via public investment and detract therefrom via reduced capital formation in the private sector. Public finances thus have an important bearing on the rate of economic growth, a fact which has been dealt with throughout the literature, and which was central to Ricardo's critique of the public sector. Moreover, the choice between tax and loan finance becomes an instrument of intertemporal burden distribution. Since loan finance falls more heavily on capital formation, it leaves future generations with a smaller capital stock. Considerations of intergenerational equity thus permit public investment, the benefits from which accrue to future generations, to be loan financed, while calling for current services to be tax financed. This establishes the rationale for a dual budget system, with balance in the current and loan finance in the capital budget. This reasoning, in turn, calls for the inclusion of depreciation in the current budget, with corresponding debt retirement over the useful life of the asset. It should be added that inclusion of outlays in the capital budget does not require public acquisition of real assets (with its fictitious analogy to a balance sheet) but simply the creation of future benefits, including these of investment in human capital through outlays on health and education.

A quite different macro-perspective on public expenditures emerged with the Keynesian model. While the classical framework had left the budget aggregate-demand neutral, deficit and surplus finance now became a source of demand expansion and restriction. With initial emphasis on fiscal expansion (restriction) directed at increase (decrease) in public spending, tax reduction (increase) subsequently entered as an alternative way of achieving similar results. The early Keynesian model, which left money impotent and viewed fiscal policy as allpowerful, was modified in the neoclassical synthesis of the 1950s and 1960s, and attention moved to the correct mix of fiscal and monetary constraint. Moreover, the supremacy of aggregate demand controls become questionable as attention shifted from full employment to inflation. Nevertheless, aggregate demand effects of fiscal operations have remained of major concern, joining the earlier issues of allocation and distribution as a third dimension of fiscal economics.

Fiscal Politics

The preceding discussion, in line with the tradition of fiscal economics, has dealt with normative issues of expenditure policy, that is, why such expenditures are needed and how they should be designed to obtain efficient results. More recently, a new perspective has been added. Not resting on the assumption that prescription of correct policy will be followed once laid down by economic analysis, attention has turned to how public policy does in fact behave and how its behaviour is determined. In the process, emphasis has shifted from concern with market failure to focus on failure in public policy (Buchanan and Tullock 1962). Early efforts to develop a theory of public-sector behaviour had been made in a Marxist framework, viewing the state as an instrument of exploitation by the ruling class. Recent analysis proceeds in analogy to microeconomics, involving the self-interested behaviour of voters, bureaucrats and politicians. An important focus in this analysis has been the growth of the public sector, the extent to which it reflects the inherent needs of modern society as expounded in Wagner's Law (1883), or a malfunctioning of the fiscal system based on an inherent bias towards over-expansion (*see* "▶ Public Choice").

Taxation

We now turn to the tax side of the fiscal picture, beginning with the normative requirements for a good tax system.

Criteria for Equity

From Adam Smith on, students of taxation have been concerned with the qualities of a good tax system. One such requirement, traditionally ranked first, is that the tax burden should be distributed in an equitable fashion. This requirement has taken two forms, one calling for taxation in line with benefits received, and the other for taxation in line with ability to pay. Both approaches were reflected in Smith's maxim that 'the subject of every state ought to contribute towards the support of the government, as nearly as possible in proportion to their respective abilities, that is, in proportion to the revenue which they respectively enjoy under the protection of the state' (Smith 1776, vol. II, p. 310). The benefit principle has the advantage that it links the tax and expenditure side of the budget and thus relates to the theory of public goods. The Lindahl price, after all, was the benefit tax par excellence. But benefits are not readily assigned, thus leaving the benefit rule inoperative in most cases. Moreover, as noted before, fee finance, related to the level of individual consumption of public goods, interferes with their efficient provision; nor does the benefit principle admit redistributional uses of the fiscal process.

The ability to pay approach in turn has the disadvantage that it views the distribution of the tax burden (or the resulting change in the distribution of the tax income) as independent of the expenditure side of the budget. Nevertheless, this approach has received primary attention. Beginning with J.S. Mill (1848, p. 804), taxation was viewed as imposing a sacrifice and the problem was how to distribute this sacrifice in an equitable fashion. Justice requires that people in equal positions be taxed equally so as to undergo an equal sacrifice, people in unequal positions, however, are to pay unequal amounts of tax, differentiated so as to involve an equal sacrifice. Underlying this formulation was the assumption of declining, similar, and comparable marginal utility of income schedules. Subsequent refinement by Edgeworth (1897) and Pigou (1928) differentiated between equal absolute, equal proportional, and equal marginal (least total) sacrifice. Equal marginal sacrifice calls for maximum progression provided only that the utility schedule is declining. Equal absolute sacrifice calls for progressive proportional or regressive taxation, depending upon whether the elasticity of the marginal utility of income schedule falls short of, equals, or exceeds unity. No simple rule, finally, can be given for the case of proportional sacrifice. Authors such as Edgeworth and Pigou, as had Bentham (1802), opted for the equal marginal sacrifice rule, given the utilitarian premise that least total sacrifice (or a maximum level of remaining welfare) should be the goal of rational conduct. Given the further assumption of equal utility schedules, the formulation calls for a move towards equal distribution. But having drawn this basic conclusion, it is then qualified to allow for incentive effects and the resulting shrinkage in the overall level of income which is available for distribution.

The sacrifice theory of tax equity nicely fitted the framework of the 'old welfare economics', which was willing to assume inter-personal utility comparison. As this assumption was discarded in the 1930s, equal sacrifice rules became inoperative but subsequently were replaced by the hypothesis of a social welfare function, assigning marginal social utilities to various levels of income. People with equal income should still pay the same tax (the principle of horizontal equity), while differential taxation at different levels of income would be determined in line with society's view of declining social utility as income rises. Notwithstanding Arrow's demonstration that a social welfare function cannot be derived in an unambiguous fashion, such a concept is now widely used in policy evaluation, including cost-benefit analysis and the setting of optimal tax rates.

Definition of Income

Dating back to Adam Smith, the analysis of tax equity has focused on income as the index of ability to pay. While expenditures or consumption have entered as alternatives, income has received the major attention. But the definition of income for purposes of taxation is not obvious, especially not in the context of a highly complex financial and industrial economy where income may be received and used in a variety of forms. A large part of tax analysis has thus been concerned with the definition of income and its application in this complex setting.

The analysis has proceeded from the basic concept of net income (Schanz 1896; Simons 1938) as accretion to wealth or, which is the same, increase in net worth plus consumption.

On this basis, a host of specific issues are dealt with, including the treatment of unrealized gains, depreciation, interest paid, income in kind, imputed income, as well as many other items. While economists have argued for a broad and comprehensive income base which would permit the needed revenue to be obtained at lower rates, they have had only limited success. The tax base has been diluted by an expanding net of tax preferences and it remains to be seen whether a political consensus for base-broadening can be reached. The problem is complicated by the fact that not all omissions from the tax base reflect gross efforts at tax avoidance. Others may be viewed as providing incentives to secure policy objectives which may be valid on their own terms. Economists have opposed such use of 'tax expenditures', noting that the underlying policies, if valid, may be pursued more efficiently through the expenditure side of the budget, including subsidies to solicit private sector responses. More recently, the problem of tax base definition has been complicated further by the impact of inflation. With ability to pay relating to real rather than nominal income, inflation adjustments are appropriate, both with regard to the indexing of rate brackets and the measurement of capital gains and interest income.

A further central issue relates to the tax treatment of the corporation. The purist position has been that equity in taxation refers to the tax treatment of individuals, and that all income ultimately belongs to them, be it directly or via their ownership of legal persons such as corporations. From this it is concluded that corporate source income should be taxed to its owners, and not be subjected to an additional or absolute tax at the corporate level. Dividend and interest disbursement by the corporation should be taxed to the shareholder, as should undistributed earnings. For purposes of administration, shareholders' taxes on corporate source income may be withheld at the corporate level, but they would then be credited under the individual income tax. This approach, however, is rarely followed. Instead, a separate corporate tax is imposed and corporate source income if retained is allowed for but imperfectly at the individual shareholder level.

Finally, and of increasing importance, the structural problems of the income tax are complicated by international capital flows and multiple jurisdictions. Techniques have been devised to protect capital income against multiple taxation, and the question of which jurisdiction (e.g. origin or destination) is entitled to a particular tax base has been debated.

Consumption Base

While primary focus has been on the nature of income as the tax base, consumption has been considered as an alternative thereto. Hobbes (1651) argued at an early point that a person should be taxed on what he takes out of the pot (i.e. consumes) and not on what he adds (i.e. saves). Also, economists from Mill to Marshall, Pigou and Fisher, have held that the income tax involves a 'double taxation' of saving. By taxing income when saved and then taxing interest thereon, the tax differentiates unfairly against savers and in favour of consumers, and an excess burden or efficiency cost results from such tax discrimination against saving. The case for a consumption base was impeded, however, by the assumption that it would have to be applied in the form of 'in rem' taxation, i.e., through excises or general growth income taxes such as the retail sales or value added tax (*see* Kay 1987). Because of their impersonal nature, such taxes would not be acceptable on equity grounds. This objection no longer applies, as the case for the consumption base has been reformulated in the context of a personal expenditure tax (Kaldor 1955), with personal exemptions and progressive rates applicable as under the income tax.

Much recent attention has been given to the way in which a comprehensive expenditure base would be computed. Determined as cash income plus net borrowing minus net investment, consumption would be arrived at as a residual, rather than by attempting the aggregation of outlays. Determination of the expenditure base would bypass certain central difficulties of income determination (especially the treatment of postponed income, unrealized gains, and depreciation), but new problems would arise as well, and pressures for base preferences would not disappear.

Difficulties would have to be met, especially in relation to the transition from income to expenditure taxation.

There is also the question whether consumption is indeed the correct base. The expenditure tax avoids disincentive to saving and gives equal treatment to individuals who consume early and late in their lives. But preference is given to those who make gifts and leave bequests. Thus critics hold that the index of equality should be defined as equal present value of potential (not only actual) consumption. It then follows that bequests and gifts should be included in the testator's expenditure base. Also, it may be argued that the gain from saving not only consists in increased future consumption but that the holding of wealth itself carries utility, allowance for which may call for a supplementary wealth tax. Viewed as the equivalent of a wage income tax, the expenditure tax stands in uneasy contrast to the traditional view that if anything income from capital should be taxed more heavily than income from wages.

The preceding discussion of the tax base has focused on income and expenditures as the primary options. In a fuller treatment, other forms of taxation, in particular the property tax and payroll tax, would have to be considered as well. Indeed, the growth of tax structures has reflected the changing patterns of economic institutions and availability of 'tax handles'. What are good taxes for a highly developed financial economy such as the US of today were not feasible when selective property taxes provided the main revenue source under colonial conditions. Nor can the same tax rules be applied to developing countries of today. Moreover, the choice of appropriate taxes differs at the central and local levels of government, all of which renders the problem of tax structure design richer and more complex than can be accounted for here.

Efficiency Rules

An equitable distribution of the tax burden is one important attribute of a good tax structure. But it is not the only one. We now turn to the further and related issue of efficiency in taxation. As Adam Smith noted (1776, p. 310), taxes ought to be designed so 'as to take out of the pockets of the people as little as possible over and above what it brings into the public treasury of the state'. Compliance and collection costs should be minimized, but this is not all. At a more subtle level, as later discussion has shown, a given revenue should be drawn from any one taxpayer so as to impose the least welfare loss. Taxes other than lump sum taxes impose an efficiency cost, i.e., leave the taxpayer with a loss which exceeds the value of revenue which government obtains. In the extreme case, a taxpayer may be burdened while there is no revenue gain: for example, a person may cease to consume a taxed product, leaving the treasury without gain and forcing the taxpayer into a less satisfactory consumption mix. Or, imposition of an income tax may induce the taxpayer to substitute leisure for income, thereby reducing the tax base while burdening him with a less satisfactory work–leisure choice.

The measurement of deadweight loss or loss of consumer surplus as a triangle under the demand curve was anticipated by Dupuit (1844) and Jenkin (1871), and was subsequently developed by Marshall (1890, Book III, Chapter 6). Modern discussion of deadweight loss begins with Pigou's treatment of announcement effects (1928). Assuming leisure to be fixed, the optimal solution (which minimizes deadweight loss) calls for all products to be taxed at a uniform ad valorem rate, but the problem is more difficult if leisure is allowed to vary. Since leisure as such cannot be taxed, the taxation of products complementary to leisure must take its place. As first shown by Ramsey (1927), deadweight loss is minimized by imposing a set of differential ad valorem rates, such as to reduce the production of all commodities in equal proportion. After an interval of nearly fifty years, this rule then laid the basis for the theory of optimal taxation (Diamond and Mirrlees 1971). Discussed elsewhere (*see* "▶ Optimal Taxation"), it will not be expanded on here.

Further problems arise in moving from the optimal treatment of a particular taxpayer to that of the group. If the utility function of all taxpayers is assumed to be the same, no difficulty arises. But if it is allowed to differ, the ideal pattern of optimal taxation would call for the tailoring of differential rates of tax to the particular preferences of each taxpayer. Since this is impossible, a general tax formula has to be used, based on representative behaviour. This bypasses issues of horizontal equity, issues which arise precisely because behaviour patterns differ.

Shifting and Incidence

Economists have for long been aware that there exists a difference between the point at which a tax is imposed (its statutory incidence) and that at which its final incidence comes to rest. The in-between process or shifting has filled the largest chapter in the history of public finance, and by its nature as market economics has developed in close linkage to the general body of economic theory.

In the context of the Physiocratic model, only a tax on land could be productive as only land was a true source of income. The classics continued to focus on the division of output among factor shares, but the addition of capital to land and labour provided a three-factor model. This expanded model not only fitted the analytical scheme but also reflected the social structure of the times. Ricardo in particular devoted a large part of his treatise to this aspect of taxation. He agreed with the Physiocrats that a tax on rent cannot be shifted, but replaced the view of land as the basic source of income by recognition of rent as an intra-marginal return which does not affect price. A tax on wages must in the short run be borne by profits, so he held, since wages are at subsistence and cannot be reduced. But accumulation declines in the longer run, forcing a reduction in population. The same holds for a tax on necessities. Taxes on luxury products are absorbed by the payee as are taxes on profits. But the latter once more reduce accumulation, and hence the demand for labour. In the end only luxury consumption remains as a solid tax base.

This simple solution crumbled with the subsistence hypothesis. Replaced by a generalized theory of factor pricing, based on marginal products, no factor share remained immune to taxation, and tax incidence had to be viewed in the context of a general equilibrium system of competitive factor and product pricing (Walras 1874). With factor and product pricing interacting in a general equilibrium setting, a product tax might come to affect the position of households from the sources as well as from the uses side of their account, just as a tax on factor income might come to be felt from the uses as well as from the source side. Advancing in many directions, the theory of incidence came to distinguish between partial and general taxes, short and long run results, as well as outcomes in competitive and imperfect markets.

Moreover, while the classics had been concerned primarily with the distribution of the burden among factor shares, subsequent concern turned to the more complex issue of burden allocation among income groups.

Analysis of partial product taxes in terms of supply and demand curves and their elasticities were first undertaken by Jenkin (1871), developed by Marshall (1890), and extended in detail by Edgeworth (1897). As concluded later (Hicks 1947), a tax would be divided between buyers and sellers in inverse relation to the elasticity of substitution in supply and demand. As first suggested by Barone (1899) and later developed by J.R. Hicks (1939), a distinction was drawn between the income and substitution effects of a tax. As the two effects contradict each other, the net effect of an income tax on factor supply was no longer evident. Moreover, it no longer followed that a progressive tax schedule must depress work effort more than would a proportional one. The former, to be sure, involves higher marginal rates at high levels of income, and hence imposes a more severe substitution effect on such taxpayers; but the latter requires higher rates further down, so that the net effect is not evident.

While most incidence analysis has been conducted in competitive markets, attention has also been given to non-competitive conditions. Cournot (1838) early on showed that a tax on monopoly profits cannot be shifted, but later analysis dealing with other forms of market imperfection showed that profits taxes may indeed be passed on. Returning to the competitive case, analysis has focused on the incidence of profits taxes imposed on selected industries only. As Marshall (1890) had pointed out, such profits in the short run are quasi-rents so that the tax stays put; but given sufficient factor mobility, such is not the case in the longer run. A reduction in the return to capital in any particular industry eventually comes to be shared by capital at large. As capital moves from the taxed to tax-free industries, net returns are equalized. In the process, consumers and other factors may come to share part of the burden (Harberger 1974).

The emergence of neoclassical growth models (Solow 1956) soon invited a reformulation of long run incidence in the classical tradition (Krzyzaniak 1967; Feldstein 1974). Incidence under steady growth is shown to depend on savings propensities as well as the elasticities of factor supplies. Thus substitution of a tax on capital income for an equal tax on labour income will leave part of the burden on labour, even if factor supplies are inelastic, provided that the propensity to save out of capital income is higher; and capital income will bear the entire burden, even if labour supply is elastic, provided that the propensities to save are the same.

As is not infrequently the case, theory advanced more rapidly than its empirical verification. Econometric testing of incidence outcomes has been undertaken but rarely (Musgrave and Krzyzaniak 1963) and has led to controversial results. Instead, two more hypothetical approaches have been undertaken towards quantitative estimation of tax-burden distribution. One approach has relied on what seem reasonable assumptions regarding the shifting of particular taxes, which assumptions are then implemented on the basis of available income and expenditure data (Colm and Tarasov 1940; Pechman and Okner 1974). The other involves simulation of a general equilibrium model, reflecting the observed structure of the economy. This model is then made to respond to the introduction of particular taxes (Shoven and Whalley 1984), and the resulting changes in household positions are observed. The former approach has the advantage that the implications of various shifting hypotheses can be tested, but it fails to allow for second round effects. The latter has the advantage of accounting for a full sequence of adjustments and includes deadweight losses in the burden estimation, but it has the disadvantage that the result are drawn from the premise of perfectly competitive markets. In comparing the two, much depends on the weight of first round effects.

See Also

- ▶ Neutral Taxation
- ▶ Optimal Taxation
- ▶ Progressive and Regressive Taxation
- ▶ Ricardian Equivalence Theorem
- ▶ Welfare Economics

Bibliography

Barone, E. 1899. Di alcuni teoremi fondamentali per la teoria matematica dall' imposta. *Giornale degli Economisti,* Second series 60. Trans. as 'About fundamental theorems on the mathematical theory of taxation' in Musgrave and Shoup (1959).

Bentham, J. 1802. Principles of the civil code. In *The theory of legislation,* ed. C. Ogden. London: Kegan, 1931.

Buchanan, J., and G. Tullock. 1962. *The calculus of consent*. Ann Arbor: University of Michigan Press.

Colm, C., and H. Tarasov. 1940. *Who pays the taxes?* Washington, DC: Temporary National Economic Committee.

Cournot, A. 1838. *Recherches sur les principes mathématiques de la théorie des richesses*. Paris: Hachette.

Diamond, P., and J. Mirrlees. 1971. Taxation and public production I: Production and efficiency; and II: Tax rules. *American Economic Review* 61(Pt I), March: 8–27; (Pt II), June: 261–278.

Dupuit, J. 1844. De la mésure de l'utilité des travaux publics. *Annales des Ponts et chausées,* Second series, vol. 8. Reprinted in *International Economic Papers No. 2* (1952): 83–110. London: Macmillan.

Edgeworth, F.Y. 1897. The pure theory of taxation. *Economic Journal* 7(Pt I), March: 46–70; (Pt II), June: 226–238; (Pt III), December: 550–571.

Feldstein, M. 1974. Social security, induced retirement, and aggregate capital accumulation. *Journal of Political Economy* 82: 905–926.

Harberger, A.C. 1974. *Taxation and welfare*. Boston: Little, Brown.

Hicks, J. 1939. *Value and capital*. Oxford: Oxford University Press.

Hicks, U. 1947. *Public finance*. New York: Pitman.

Hobbes, T. 1651. *Leviathan*. Harmondsworth: Penguin Books, 1968.

Hume, D. 1739. *A treatise of human nature*, ed. L. Selby-Bigge. Oxford: Oxford University Press, 1975.

Jenkin, F. 1871. On the principles which regulate the incidence of taxes. *Proceedings of the Royal Society of Edinburgh.* Reprinted in Musgrave and Shoup (1959).

Kaldor, N. 1955. *An expenditure tax*. London: Allen & Unwin.

Kay, J. 1987. Indirect taxes. In *The new palgrave: A dictionary of economics*, ed. J. Eatwell, M. Milgate, and P. Newman, vol. 2. London: Macmillan.

Krzyzaniak, M. 1967. The long-run burden of a general tax on profits in a neoclassical world. *Public Finance* 22: 472–491.

Lindahl, E. 1919. *Die Gerechtigkeit der Besteuerung*. Lund: Gleerup.

Marshall, A. 1890. *Principles of economics*, 8th ed. London: Macmillan, 1920.

Mazzola, U. 1890. *I dati scientifici della finanza pubblica*. Rome: E. Loescher & Co.

Mill, J.S. 1848. *Principles of political economy*, ed. W. Ashley. London: Longmans, 1921.

Musgrave, R.A. 1939. The voluntary exchange theory of taxation. *Quarterly Journal of Economics* 53: 213–237.

Musgrave, R.A. 1959. *The theory of public finance*. New York: McGraw-Hill.

Musgrave, R.A., and M. Krzyzaniak. 1963. *The shifting of the corporation tax*. Baltimore: Johns Hopkins Press.

Musgrave, R.A., and C.S. Shoup, eds. 1959. *Readings in the economics of taxation*. Homewood: Irwin.

Pechman, J., and B. Okner. 1974. *Who bears the tax burden?* Washington, DC: Brookings Institution.

Pigou, A. 1920. *The economics of welfare*. London: Macmillan.

Pigou, A. 1928. *A study in public finance*. London: Macmillan.

Ramsey, F.P. 1927. A contribution to the theory of taxation. *Economic Journal* 37: 47–61.

Ricardo, D. 1817. *The principles of political economy and taxation*, ed. P. Sraffa. Cambridge: Cambridge University Press, 1951.

Samuelson, P.A. 1954. The pure theory of public expenditure. *Review of Economics and Statistics* 36: 387–389.

Sax, E. 1887. *Grundlegung der theoretischen Staatswirtschaft*. Vienna: Holder.

Schanz, G. 1896. Der Einkommensbegriff und die Einkommenssteuergesetze. *Finanzarchiv* 13: 1–87.

Shoven, J., and J. Whalley. 1984. Applied general-equilibrium models of taxation and international trade. *Journal of Economic Literature* 22: 1007–1051.

Simons, H. 1938. *Personal income taxation*. Chicago: University of Chicago Press.

Smith, A. 1776. *An inquiry into the nature and causes of the wealth of nations*, ed. E. Cannan. New York: Putnam, 1904.

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

Tiebout, C.M. 1956. A pure theory of local expenditures. *Journal of Political Economy* 64: 416–424.

Wagner, A. 1883. Lehr und Handbuch der Politischen Oekonomie, Vierte Hauptabeilung: Finanzwissenschaft, 3rd ed. Leipzig: Winter.

Walras, L. 1874–7. *Elements of pure economics.* Trans. W. Jaffé. Homewood: Irwin, 1954.

Wicksell, K. 1896. *Finanztheoretische Untersuchungen nebst Darstellung und Kritik des Steuerwesens Schwedens*. Jena: Fischer.

Public Goods

Agnar Sandmo

Abstract

This article provides a mathematical and diagrammatic exposition of the theory of public goods as originally formulated by Paul Samuelson. It describes the extension of the model to take account of the costs of distortionary taxation, and discusses the concept of the marginal cost of public funds. Different types of public goods (such as mixed goods and local and global public goods) are discussed before a survey of the incentive problems related to preference revelation.

Keywords

Benefit theory of taxation; Clubs; Cost-benefit analysis; Direct democracy; Distortionary taxation; Externalities; Free-rider problem; Global public goods; International migration; Lindahl equilibrium; Local public finance; Local public goods; Lump-sum taxes; Marginal cost of public funds; Marginal rate of substitution; Marginal rate of transformation; Market failure; Mixed goods; Non-rivalry in consumption; Optimal taxation; Pigou, A. C.; Public factors of production; Public finance; Public goods; Pure public goods; Redistribution of income and wealth; Revealed preferences; Samuelson, P. A.; Second best; Shadow pricing; Single-peaked preferences; Tax incidence; Tiebout hypothesis; Voluntary provision of public goods; Voting; Willingness to pay

JEL Classifications

H4

The development by Paul Samuelson (1954, 1955) of the modern theory of public goods must be counted as one of the major breakthroughs in the theory of public finance. In two very short papers Samuelson posed and partly solved the central problems in the normative theory of public expenditure:

(a) How can one define analytically goods that are consumed collectively, that is, for which there is no meaningful distinction between individual and collective consumption?
(b) How can one characterize an optimal allocation of resources to the production of such goods?
(c) What can be said about the design of an efficient and just tax system which will finance the expenditures of the public sector?

None of these questions was entirely new to the literature of public finance. Indeed, more than 250 years ago David Hume (1739) noted that there were tasks which, although unprofitable to perform for any single individual, would yet be profitable for society as a whole, and which could therefore only be performed through collective action. The theme was later taken up by Hume's friend Adam Smith, who maintained that one of the duties of the state consisted in

> erecting and maintaining certain publick works and certain publick institutions, which it can never be for the interest of any individual or small number of individuals, to erect and maintain; because the profit would never repay the expense to any individual or small number of individuals, though it may frequently do much more than repay it to a great society. (Smith 1776, pp. 687–8)

Apart from this insight, however, the progress made over the next centuries, certainly with regard to problems (a) and (*b*), was rather modest. From the point of view of the history of ideas, this is hardly surprising. What is required is a satisfactory theory of market failure. But this presupposes a clear understanding of the optimality properties of the market allocation of resources, which was not established until the modern development of Paretian welfare economics in the late 1930s. More was undoubtedly achieved with respect to problem (*c*), reflecting the fact that problems of tax incidence had been a central area of theoretical analysis ever since the time of the classical economists, and that criteria of just taxation had developed independently of any analysis of the expenditure side of the public budget. Still, Samuelson's formulation was in every respect a great leap forward, presenting an integrated solution to all three problems, and determining the research agenda for the years to come. It is therefore natural to begin by setting out the basic elements of his model.

In a short article it is of course impossible to do justice to the large literature in this field. For more comprehensive surveys the reader is referred to the textbooks by Atkinson and Stiglitz (1980, lectures 16–17) and Myles (1995, ch. 9), and the article by Oakland (1987).

The Samuelson Model

The aim of the model is to derive conditions for optimal resource allocation in an economy in which there are two types of goods, private and public. It is worth emphasizing that these terms do not prejudge the respective tasks of the private and public sectors; the analysis at this stage is institution-free and can best be considered as representing the problems of a planner who knows the production possibilities of the economy, the preferences of the consumers and his or her own ethical values. The definition of the two types of goods is technological, not institutional.

The nature of the two types of goods is defined by the equations which give the relationship between individual and aggregate consumption. For private goods the total quantity consumed is equal to the sum of the quantities consumed by the individuals, so that

$$x_j = \sum_{i=1}^{I} x_j^i, (j = 0, \ldots, J) \qquad (1)$$

where the superscript refers to individuals and the subscript to commodities. For public goods the corresponding relationship is one of *equality* between individual and total consumption, namely

$$x_k = x^i_k, (i = 1, \ldots, I; k = J + 1, \ldots, J + K). \tag{2}$$

Individual preferences, represented by utility functions, are then defined over the quantities consumed of private and public goods, so that we can write the utility of individual *i* as

$$\begin{aligned} U^i &= U^i\left(x^i_0, \ldots, x^i_J, x^i_{J+1}, \ldots, x^i_{J+K}\right) \\ &= U^i\left(x^i_0, \ldots, x^i_J, x_{J+1}, \ldots, x_{J+K}\right), (i = 1, \ldots, I). \end{aligned} \tag{3}$$

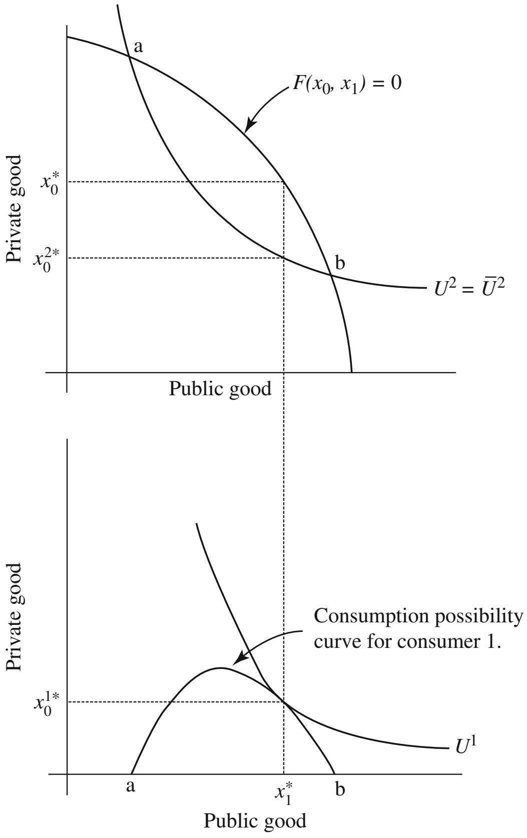
The definition (2) has given rise to some confusion and controversy. Are there actually any goods which can be described by this definition? The usual answer is that there are some cases of 'pure' public goods, like national defence, which can indeed be so described; in such cases consumer benefits are directly related to the total availability of the good in question, and the consumption benefits of any one individual do not depend on the benefits enjoyed by others. This property of public goods is usually referred to as non-rivalry in consumption; given the supply of the good in question, the consumption possibilities of one individual do not depend on the quantities consumed by others as they do in the case of private goods. However, many goods that one naturally thinks of as public turn out on closer inspection to have elements of rivalry. A road may satisfy the definition of a public good as long as volume of traffic is low, but with higher density and consequent congestion this will no longer be the case. Accordingly, several studies have been devoted to the analysis of 'impure' public goods, combining in some way the properties of private and public goods in the original Samuelson definition; we shall return to this below.

It should be observed, however, that the Samuelson formulation does not assume that the *benefits* derived from the supply of the public good are the same for all, even though a*vailabilities* are the same. Neither does it assume that the benefits from public goods are independent of the quantities consumed of private goods. And the elements of rivalry in the road congestion example may be captured by introducing externalities in the consumption of a private good – car use – whose benefits depend on the supply of a public good – the road. Thus, the original Samuelson formulation offers great flexibility of interpretation, and we have been provided with an answer to the first of the main problems noted above.

We now turn to the problem of optimality of resource allocation and begin by characterizing a Pareto optimum for this kind of economy. Since the interesting special features of the model are on the consumption side only, we assume that the conditions for efficient production are satisfied, so that the production possibilities for the economy can be summarized in the transformation or production possibility equation

$$F(x_0, \ldots, x_J, x_{J+1}, \ldots, x_{J+K}) = 0. \tag{4}$$

The problems of Pareto optimality may now be formulated as follows: of all allocations satisfying Eq. 4, find the allocation which maximizes utility for consumer 1, given arbitrary but feasible utility levels for all other consumers. As shown by Samuelson (1955), the solution can be given an instructive graphical solution in the two-dimensional case. We therefore begin with the case where there are two consumers and one private and one public good. In the upper panel of Fig. 1 we have drawn the production possibility curve as well as an indifference curve corresponding to the fixed level of utility for consumer 2; since the two curves intersect, there are obviously a number of allocations which satisfy these two constraints. In the lower panel the curve as shows the consumption possibilities for consumer 1, the points a and b corresponding to the points of intersection in the upper panel. For any point on U^2 between a and b, it must be the case that the two individuals consume the same amount of the public good, while consumer 1's private good consumption is equal to the vertical difference between the production possibility curve and consumer 2's indifference curve. The best allocation from 1's point of view is then given by the



Public Goods, Fig. 1 Pareto optimality with one private and one public good

tangency between his indifference curve and the consumption possibility curve in the lower panel. This determines the optimum supply of the public good $\left(x_1^*\right)$ and consumer 1's consumption of the private good $\left(x_0^{1*}\right)$ as well as the consumption of consumer 2 $\left(x_0^{2*}\right)$.

The slope of the consumption possibility curve must of course be equal to the difference of the slopes of the two curves from which it is derived. The tangency point can therefore be characterized in terms of marginal rates of substitution and transformation as

$$\mathrm{MRS}^1 = \mathrm{MRT} - \mathrm{MRS}^2, \text{or } \mathrm{MRS}^1 + \mathrm{MRS}^2 = \mathrm{MRT}.$$

In more precise mathematical terms this condition can be rewritten (if we let subscripts denote partial derivatives) as

$$\frac{U_1^1}{U_0^1} + \frac{U_1^2}{U_0^2} = \frac{F_1}{F_0}. \qquad (5)$$

In words: the sum of the marginal rates of substitution should be equal to the marginal rate of transformation between the public and the private good. Or, since the private good may be taken as a numeraire commodity, the sum of the marginal willingness to pay for the public good should be equal to the marginal cost of production. The intuition should be clear: an extra unit of supply benefits both consumers simultaneously; to find the total marginal benefit we have to take the sum of the marginal benefits accruing to all consumers. Problem (2) has been solved.

The mathematical derivation of the corresponding condition in the general case need not occupy us here. To extend the analysis to more than two consumers, we have only to add more terms on the left-hand side of (5). An increase in the number of public goods simply requires us to introduce similar conditions for every such good. To generalize to an arbitrary number of private goods, we note that for any given allocation of public goods, the allocation of private goods should be a Pareto optimum relative to this, so that the usual marginal conditions must hold. This gives us two sets of first order conditions for Pareto optimality, namely:

$$\frac{U_j^i}{U_0^i} = \frac{F_j}{F_0}, \ (i = 1, \ldots, I, J = 1, \ldots, J) \qquad (6)$$

$$\sum_{i=1}^{I} \frac{U_k^i}{U_0^i} = \frac{F_k}{F_0}, \ (k = J+1, \ldots, J+K). \qquad (7)$$

In the two-dimensional case the first order conditions could be taken to describe a true maximum because the diagrams introduced the required convexity–concavity conditions. In the more general case one has to assume quasi-concavity of the utility functions as well as convexity of the transformation surface for the second order conditions to be satisfied.

There is of course an element of arbitrariness in the concept of Pareto optimality, corresponding to the arbitrary location of consumer 2's indifference curve in Fig. 1. The model can be closed by assuming the existence of a social welfare function, and the usual assumption is that this is of the Bergson-Samuelson type, where the arguments of the function are the individual utility levels. Maximizing the welfare function $W(U^1, \ldots, U^I)$ gives as the optimality conditions first (6) and (7) – since a welfare optimum must be a Pareto optimum – and then a set of conditions for optimal distribution of consumption between individuals. These can be written as

$$W_i U_0^i = W_h U_0^h, \; (i, h = 1, \ldots, I). \qquad (8)$$

The marginal social utility of consumption should be the same for all. (Note that although the conditions as stated here refer to the consumption of private good 0, they can be converted, by using conditions (6), to express the equality of the marginal social utility of consumption in terms of any private good.)

Suppose now that private goods are allocated through a system of perfectly competitive markets, and that the allocation of resources to public goods also satisfies the efficiency conditions (7) as the result of some decision procedure which is yet to be specified. Imagine further that at least part of the provision of public goods is undertaken by the public sector, and that taxes are needed to finance this. What is the ideal tax system for this purpose? We wish the tax system to satisfy conditions (8), but these are conditional on the remaining first-order conditions being satisfied. Under competitive conditions the marginal rates of substitution will be equal to consumer prices, if we take commodity 0 to be the numeraire good, while marginal rates of transformation will correspond to producer prices. Thus, conditions (6) will be satisfied in a competitive economy provided that consumer prices are equal to producer prices. But this means that there must be no distortionary taxation; the only taxes that are consistent with a fully optimal solution are lump-sum taxes in amounts independent of all components of demand and supply for consumers and firms. This insight is of course well known from the standard competitive model with private goods only, but it is worth restating in the present context as the answer to problem (*c*).

This exposition of the basic elements of the Samuelson model can be used to put his contribution into historical perspective. Earlier writers on public finance, for example Mazzola (1890), Sax (1924) and Pigou (1928), did in fact apply marginal utility theory to the problem of the optimal supply of public goods, emphasizing the optimality rule that marginal benefit at the optimum should be equal to marginal cost. They failed, however, to develop a definition of public goods that could be used to characterize the difference between such goods and private goods. For the same reason they were also vague about the nature of the marginal benefit and how to measure it in the absence of market prices. Finally, although there is much interesting discussion by the older writers of the ability to pay and benefit theories of taxation, the efficiency aspect of taxation played a very minor part in their writings, and so they were unable to face the basic problem of how to reconcile the objectives of a just distribution and economic efficiency. With the Samuelson formulation all these issues had been clarified, and the foundation had been laid for further progress.

Distortionary Taxation

The above optimality rules hold for the case where taxation is non-distortionary, that is, where taxes are imposed to raise revenue and to redistribute incomes without disturbing the efficiency properties of the price mechanism. For a variety of reasons such taxes are hardly feasible, and it is interesting to consider the modifications that will have to be made if taxes are distortionary. Pigou (1928) argued that the cost of tax distortions should be taken into account in balancing the costs and benefits of public goods supply:

> Where there is indirect damage, it ought to be added to the direct loss of satisfaction involved in the withdrawal of the marginal unit of resources by taxation, before this is balanced against the satisfaction yielded by the marginal expenditure. (Pigou 1928, p. 34)

As pointed out by Atkinson and Stem (1974), however, this argument is not necessarily correct. Their analysis is an interesting exercise in the theory of the second best.

To abstract from problems of redistribution, consider the case where all individuals are identical. There are two private goods, numbered 0 and 1, and one public good, identified as commodity 2. The representative consumer maximizes his or her utility function $U(x_0, x_1, x_2)$ subject to the budget constraint

$$x_0 + P_1 x_1 = 0. \tag{9}$$

Thus, there is no lump-sum income, and commodity 0 serves as the numeraire.

Given the optimum of the consumer, the government maximizes the sum of the utility functions (a special case of the welfare function in the previous section) subject to the constraint that the resource cost of public goods supply equals the tax revenue. Thus, the government maximizes $IU(x_0, x_{1;}\, x_2)$ subject to

$$It_1 x_1 = p_2 x_2. \tag{10}$$

Here I is as before the number of consumers, t_1 is the tax per unit of commodity 1 such that $P_1 = p_1 + t_1$. The small *p's* denote producer prices, which for convenience are taken to be constant, corresponding to constant unit costs of production in terms of the numeraire. The government determines t_1 and x_2 simultaneously.

The analytical details of the model need not concern us here. To understand the result, one should note that from the formulation of the consumer's problem it follows that demand for the taxed good depends on the supply of the public good, so that the demand function can be written as $x_1 = x_1\,(P_1, x_2)$. Thus, when the supply of the public good is increased, there will be two effects on the demand for private goods. One is the effect via increased availability of the public good, another is the price effect via increased taxation. It can be shown that the condition corresponding to the Samuelson Eq. 7 in this case becomes

$$\sum_i \mathrm{MRS}^i = \frac{p_2 - t_1 I(\partial x1/\partial x_2)}{1 + (t_1/x_1)(\partial x_1/\partial t_1)}. \tag{11}$$

If there is no distortionary taxation, the right-hand side becomes simply p_2, which is the marginal rate of transformation, and we are back to the original Samuelson case. An increase in the tax rate lowers the demand for the taxed good, and the corresponding term in the denominator shows that this 'blows up' the cost of the public good; this is the effect alluded to by Pigou. On the other hand, the additional term in the numerator can in principle be of either sign and may therefore reverse Pigou's conclusions. Suppose that $\partial x_1/\partial x_2$ is positive, meaning that increased supply of the public good increases the demand for the taxed good. Then the relevant social marginal cost of the public good may in fact be lower than the pure resource cost. The point is that in this case the effect of the public good on the demand for the private good serves to counteract the tax effect. The commodity tax is distortionary because it lowers consumption and production of the taxed good. If an increase of the amount of the public good serves to push the quantity of the taxed good back towards its first best optimal level, this could lower the economic cost of production.

This analysis has inspired a considerable literature about the concept of the *marginal cost of public funds* (MCF). To start from the insight provided by the formula (11), it has been suggested that practical calculations of the optimal amount of public expenditure should be based on the formula

$$\sum \mathrm{MRS}^i = \mathrm{MCF} \cdot \mathrm{MC},$$

where MC corresponds to p_2 and the presumption is that MCF >1. The use of the MCF for practical cost-benefit analysis of public goods provision – one of the more important applications of the pure theory of public goods – would therefore tend to depress the provision of public goods below the level indicated by the Samuelson rule.

This conclusion may be disputed, however. First, it is not clear that Eq. 11 supports the hypothesis that the marginal cost of public funds exceeds 1. Even if we assume, which seems reasonable, that the tax elasticity is negative, complementarity between private and public goods ($\partial x_1 / \partial x_2 > 0$) might lead the right-hand side of (11) to become less than p_2. However, since the sign and magnitude of the complementarity term must be expected to differ between different types of public sector projects, there is a good case for considering this term to be project specific and therefore not to include it in a general measure of the cost of distortionary tax finance. In this view, it is the tax elasticity of demand that is important for the MCF.

Second, there is one feature of the Atkinson-Stern analysis which calls for particular caution in practical application. This is the assumption that the government optimizes with respect to both public goods supply and the tax rate. In principle, therefore, their results are valid only for an optimal tax system, although it can be shown that the formal expression for the MCF is the same also for a non-optimal tax system (see, for example, Sandmo 1998). More importantly, however, in the more realistic case where there are many tax rates which have not been chosen optimally, there is no reason to expect that the MCF will be the same for all sources of tax finance. It will therefore be misleading to speak about the marginal cost of public funds, as if it were a general characteristic of the whole complex system of direct and indirect tax rates.

Third, in order to focus on the efficiency aspects of the problem, Atkinson and Stern made the assumption that all consumers are identical. But one of the reasons why we have distortionary taxation is that they are not, and that governments try to achieve some measure of redistribution through the design of the tax system. As shown by Sandmo (1998), an explicit modelling that takes account of the redistributive objective leads to a measure of the marginal cost of public funds where the efficiency loss from taxation may, depending on the distributional preferences embedded in the government's policies, be partly or wholly offset by distributive gains.

Types of Public Good

In line with the original Samuelson formulation we have so far limited the discussion to pure public consumption goods. Various alternative formulations have been discussed in the literature, and we shall briefly discuss some of these.

We have already observed that many consumption goods that may be classified as public turn out also to have important elements of 'privateness'. This has two aspects. In the first case it may be argued that a public good like a national park cannot really be enjoyed by the individual without expenditure on private goods such as hiking equipment, and that even such an apparently clear case of a public good should be analysed as a mixed case of a private and a public good. To some extent this argument is based on a misunderstanding of the theory. There is no presumption that the benefit that an individual derives from the availability of a public good be independent of his or her consumption of private goods. Still, it may sometimes be useful to model the interaction between private and public goods consumption in a more explicit manner than is done in the standard formulation. One way this can be done is to take as the point of departure the consumption technology approach and assume that there are some final goods such as road trips and nature hikes that are intrinsically private but that are produced by the individual consumer by means of private and public goods inputs. The second aspect of mixed goods is that the benefits enjoyed by any one individual may depend on the consumption of others, as in the cases of a crowded road or a congested national park. This aspect, too, may be handled by the consumption technology approach by letting other people's consumption of complementary private goods enter every individual's production function for the final good in question. This would be a special case of the Samuelson formulation when in addition it is assumed that some private goods create externalities in consumption. Thus, the advantage of the consumption technology approach to the theory of public goods lies not in greater generality, but in a formulation that captures in a more intuitive fashion a natural way of thinking about public goods. An additional advantage is that the theory becomes more closely related to the practice of cost-benefit analysis, where willingness to pay is typically computed not by observing preferences directly, but by calculating the private cost reductions that would follow from an increase in the provision of a public good. The theory is further elaborated in Sandmo (1973); for an alternative formulation of similar ideas see Bradford and Hildebrandt (1977).

Not all public goods are naturally analysed as consumption goods. One of the classical examples, the lighthouse, is more easily interpreted as a producer good or a factor of production. Public factors of production were first introduced into the theoretical literature by Kaizuka (1965), who derives the efficiency conditions analogous to Samuelson's for the production case. Sandmo (1972) shows how the formulation can be used to derive shadow prices for such goods when the private sector is competitive.

The Samuelson formulation implies that the availability of any public good is the same for all individuals and independent of their decisions about private goods consumption – although, as we have noted, the benefit is not. This ignores the fact that many public goods are available only to individuals residing in a particular location, and that an individual may therefore select the amount available of the public good by changing his or her place of residence. This was first pointed out by Tiebout (1956) in a paper which has since given rise to a rich literature on the important topic of local public goods and, more generally, local public finance. We shall return below to the demand-revealing aspects of mobility between communities. But it is worth noting here that, although the original application of the basic idea was to individual choice among residential communities, there are possibilities of application to other interesting areas as well. In the labour market, workers' choice among firms might be affected by public good aspects of the working environment which are specific to the individual firm. Following Buchanan (1965), 'clubs' has become the generic term for voluntary associations of individuals whose purpose is to provide the members with a public good. Internationally, country-specific public goods might influence the pattern of international migration; in this perspective, almost all public goods would be local, and the original formulation becomes a special case characterized by geographical immobility of the population. For surveys of the theory of clubs and local public goods the reader is referred to Rubinfeld (1987) and Scotchmer (2002).

At the other end of the scale from local public goods are global public goods, goods that provide benefits to the whole of the world's population. Examples of such goods are international security, global environmental quality and scientific knowledge. One might perhaps think that in this case the theory is directly applicable, since the complications associated with geographical mobility are ruled out by assumption. On the other hand, additional problems arise because the world is not one jurisdiction but composed of a number of independent nation-states. In the original Samuelson formulation, the economy is at its production possibility frontier; this is evidently a strong assumption even for a national economy, and it becomes even more unrealistic when applied to the world as a whole. Moreover, Samuelson assumed redistribution in the form of individualized lump-sum taxes and transfers; this also is an assumption which is much farther from reality when considered in a global context. Even the assumption of redistribution via progressive taxation, which is a more realistic description of national redistribution policy, is far from the economic realities of the international community of countries.

It can be shown that the problems of global production efficiency and redistribution are in fact interrelated, as one would in fact expect on the basis of the theory of the second best; see Sandmo (2003). If one takes the viewpoint of global welfare maximization and assumes that there are perfect lump-sum transfers both within and between countries, the Samuelson optimality conditions must hold for the world as a whole. In particular, there will be global production efficiency, and the social marginal utility of income must be the same for all individuals. However, if for some reason the international transfers are not made, then production efficiency is in general not desirable. If one assumes that the global welfare function displays inequality aversion, poor countries should not be required to contribute as much to the production of global public goods as their comparative advantage would otherwise call for. But the model also points to a serious problem of incentives, because each country, in deciding how much to contribute to the production of global public goods, finds itself in a strategic situation similar to that of the single individual in the nation state, who has an incentive to be a free rider on the contributions of others (see below). At least if one assumes that national governments are motivated by a fairly narrow concept of national self-interest, there is likely to be an under-supply of global public goods.

Equilibria with Public Goods

We have concentrated on the theory of public goods as an extension of welfare economics; the central question has been how to characterize optimal or efficient allocations in economies with public goods. But just as in the case of private goods it is interesting to go on from there to consider the equilibrium allocations that would follow from particular institutional arrangements in the economy and to compare these with the optimality conditions. Thus, the theory of public goods ought to be positive as well as normative, a view emphasized strongly in the influential contributions by Buchanan (for example, 1968).

The first clear formulation of a theory of public expenditure which can be given a positive interpretation was presented by Erik Lindahl (1919), who in turn was inspired by Wicksell (1896); an important modem exposition is that of Johansen (1963). In this formulation, individuals bargain over the level of public goods supply simultaneously with the distribution of the cost between them. The bargaining equilibrium is Pareto optimal, implying that the efficiency conditions (7) are satisfied. In addition, each individual pays a price in terms of private goods which is equal to his or her marginal willingness to pay. Formally, let π^i_{J+k} be the price which individual i pays for public good k, and let p_{J+k} be the producer price or marginal cost. Then the Lindahl equilibrium will be characterized by the condition

$$\sum_i \pi^i_{J+k} = p_{J+k}, (k = 1, \ldots, K). \tag{12}$$

Thus, at first glance the concept of a Lindahl equilibrium seems to establish an analogue to competitive markets for private goods with the interesting difference that prices should differ from one individual to another, depending on their marginal willingness to pay. This also ties in with older notions of the benefit theory of taxation, according to which taxes were seen as payments for public goods, to be levied in accordance with the benefits which each individual derived from them.

At the technical level it may be noted that there is an interesting ‘duality’ between the definitions of private and public goods on the one hand and the properties of equilibrium prices on the other. In terms of quantities, for private goods the sum of individual quantities consumed a*dds up* to the quantity produced, while for public goods individual consumption *equals* aggregate production. In terms of prices, on the other hand, for private goods each consumer price *equals* the producer price, while for public goods individualized consumer prices a*dd up* to the producer price.

There is, however, one crucial difference between a Lindahl equilibrium and a competitive equilibrium for private goods. With private goods, individuals facing given prices have clear incentives to reveal their true preferences by equating their marginal rates of substitution to relative prices. Without paying, the individual is excluded from enjoying the benefits of consumption. With public goods this no longer holds. Because individuals have the same quantity of public goods available to them whether they pay or not, they have an incentive to misrepresent their preferences and be free-riders on the supply paid for by others. Moreover, this problem is likely to be particularly severe when the number of individuals is large, since an individual contribution will then make little difference to the total supply. The connection between Lindahl equilibria and the game theoretic concept of the core was discussed by Foley (1970); see also the survey by Milleron (1972).

The equilibrium of the Lindahl model is not compatible with individual incentives to reveal preferences truthfully; for this reason Samuelson (1969) has referred to the individual Lindahl prices as pseudo-prices and to the equilibrium as a pseudo-equilibrium. In this case one would conjecture that, because all individuals have the same incentives to understate their true marginal willingness to pay, the Lindahl mechanism would result in equilibrium levels of public goods supply which would be too low relative to the optimum. But there is really no need to associate the problem of preference revelation with this procedure alone; as another extreme, one might think of the case where individuals are asked to state their preferences on the assumption that the cost to them is completely independent of their stated willingness to pay, but there is a positive association between this and the quantity supplied. Then there will be incentives to exaggerate the willingness to pay and a consequent tendency towards oversupply. Thus, the general problem which arises is how to design a mechanism that will allow the decision-maker to implement the efficiency condition.

Various solutions to this problem have been discussed in the literature. The most practically oriented solution is that of cost-benefit analysis, which takes as its point of departure that people's preferences for public goods are revealed in the market through their demands for complementary private goods (see above). But in theoretical terms it has been shown that this will be true only on certain rather restrictive assumptions about technology and preferences. Another solution is represented in the literature on local public goods, where it has been suggested that people reveal their preferences for public goods by moving to the community offering them their most preferred combination of taxes and public goods. But whether this process will result in an optimum satisfying the efficiency conditions must clearly depend first on how the supply of public goods is determined within each community and second on whether there are enough communities to satisfy the variations of preferences in the population as a whole. Thus, in general, observation neither of the consumption of private goods nor of individuals' mobility between local communities provides reliable information on preferences.

Presumably as a response to the problem of market failure, decisions on public goods supply are largely made by political processes. In a democracy, the natural decision-making process to study is that of voting, and there is by now a substantial literature on this. Most of this is concerned with the stylized situation where public goods supply is determined by majority voting with the consumers themselves being the voters; thus, 'direct democracy' is assumed. The first paper in this area was that of Bowen (1943), who also considered the question of when a voting equilibrium would be Pareto optimal. Later contributions have emphasized that very restrictive assumptions on preferences are sometimes required for a voting equilibrium to exist, and these – like the so-called single-peakedness assumption – are not always attractive in the public goods context. Nevertheless, voting models have become quite popular in descriptive analyses of public goods decisions, particularly at the local government level.

There has also been a great deal of interest in studying planning procedures whereby individuals find it in their own interest to reveal truthfully their preferences for public goods. The first discussion of such a procedure – although in a somewhat different context – was that of Vickrey (1961), but the more recent developments are based on the work of Clarke (1971) and Groves (1973). It is shown there that truthful preference revelation will result if individuals pay a tax on the marginal unit demanded of the public good which is equal to the difference between the marginal cost and the sum of the marginal benefits received by all other individuals. These procedures are of great theoretical interest, perhaps mainly because they clarify the nature of the free-rider problem. However, at present they seem rather far from the state where they could be implemented in practical situations; they would probably be administratively costly to operate, and they also make heavy demands on individual consumers' ability to understand and participate in the process. For surveys of this area see Tulkens (1978) and Laffont (1987).

Doubts have occasionally been voiced on whether the free-rider problem has been given too much prominence in the theoretical literature. Johansen (1977) has argued that there is no clear evidence that this is seen as a major problem in practical public sector decision-making, and suggests that individuals are much more likely to reveal their true willingness to pay than the literature indicates. This is so, he argues, both because truthfulness is a strong social norm and because it is a simple strategy that does not rely on complicated strategic considerations. There is also some empirical evidence from experimental situations to suggest that the revealed willingness to pay is not very sensitive to the associated method of cost distribution; see Bohm (1972).

The point of view taken in most of the literature considered here is that the incentive revelation problem requires decisions on public goods supply to be taken by some governmental body. However, starting with Olson (1965), there has emerged a literature on the voluntary provision of public goods. This literature is perhaps most naturally interpreted as concerned with relatively small groups, in which the incentive to free ride is limited, and not with public goods provision on a national scale. In the framework of this theory, as formulated for example by Bergstrom et al. (1986), the decision to contribute to a public good is formulated in the standard framework of consumer demand theory. Consumers allocate their incomes between private goods and contributions to public goods, which are made under assumption that the contributions of all other consumers are taken as given, and one can then study the properties of the resulting Nash equilibrium. Particular attention has been given to the effect on contributions of a redistribution of income; as first shown by Warr (1983), under some assumptions this will change individual contributions in such a way that the aggregate supply of the public good is unaffected.

Perspectives

The Samuelson theory of public goods has been of decisive influence for the theory of public expenditure, which was developed in a number of directions during the second half of the 20th century. The extensions and reinterpretations of the original theory to the cases of public factors of production, mixed (public-private) goods and local and global public goods have significantly increased the applicability of the theory. Much has also been achieved to enrich our understanding of the incentive problems that arise in actual allocation mechanisms for public goods supply; however, it is probably fair to say that the normative theory of public goods has become much more satisfactory from a theoretical point of view than the positive theory. This state of affairs may in fact be unavoidable. The normative theory has little need to model institutional details and can thus be given a more unified appearance. A positive theory, on the other hand, must to a greater extent model economic and political institutions, and there is no single institution corresponding to the competitive market in the private goods case which can serve as a unifying benchmark for the analysis. Moreover, development of the positive theory of public goods must necessarily be closely tied to the progress of the positive theory of public sector behaviour in general; it will be interesting to see whether this theory can be developed to provide descriptive models of public goods provision that are both realistic and reasonably simple.

See Also

- ▶ Incentive Compatibility
- ▶ Lindahl Equilibrium
- ▶ Public Choice
- ▶ Public Goods Experiments

Bibliography

Atkinson, A.B.., and N. Stern. 1974. Pigou, taxation and public goods. *Review of Economic Studies* 41: 119–128.

Atkinson, A.B.., and J.E. Stiglitz. 1980. *Lectures on public economics*. New York: McGraw-Hill.

Auerbach, A.J., and M. Feldstein, eds. 1987. *Handbook of public economics*. Vol. 2. Amsterdam: North-Holland.

Bergstrom, T., L. Blume, and H. Varian. 1986. On the private provision of public goods. *Journal of Public Economics* 29: 25–49.

Bohm, P. 1972. Estimating demand for public goods: An experiment. *European Economic Review* 3: 111–130.

Bowen, H.R. 1943. The interpretation of voting in the allocation of economic resources. *Quarterly Journal of Economics* 58: 27–48.

Bradford, D.F., and G.G. Hildebrandt. 1977. Observable preferences for public goods. *Journal of Public Economics* 8: 111–131.

Buchanan, J.M. 1965. An economic theory of clubs. *Economica* 33: 1–14.

Buchanan, J.M. 1968. *The demand and supply of public goods*. Chicago: Rand McNally.

Clarke, E.H. 1971. Multipart pricing of public goods. *Public Choice* 11: 17–33.

Foley, D.K. 1970. Lindahl's solution and the core of an economy with public goods. *Econometrica* 38: 66–72.

Groves, T. 1973. Incentives in teams. *Econometrica* 41: 617–631.

Hume, D. 1739. *A treatise of human nature*. London: J. Noon.

Johansen, L. 1963. Some notes on the Lindahl theory of determination of public expenditure. *International Economic Review* 4: 346–358.

Johansen, L. 1977. The theory of public goods: Misplaced emphasis? *Journal of Public Economics* 7: 147–152.

Kaizuka, K. 1965. Public goods and decentralization of production. *Review of Economics and Statistics* 47: 118–120.

Laffont, J.-J. 1987. Incentives and the allocation of public goods. In Auerbach and Feldstein (1987).

Lindahl, E. 1919. *Die Gerechtigkeit der Besteuerung*. Lund: Gleerup. Partial translation as 'Just taxation: A positive solution' in Musgrave and Peacock (1958).

Mazzola, U. 1890. The formation of prices of public goods. In Musgrave and Peacock (1958).

Milleron, J.-C. 1972. Theory of value with public goods: A survey article. *Journal of Economic Theory* 5: 419–477.

Musgrave, R.A., and A.T. Peacock, eds. 1958. *Classics in the theory of public finance*. London: Macmillan.

Myles, G.D. 1995. *Public economics*. Cambridge: Cambridge University Press.

Oakland, W.H. 1987. Theory of public goods. In Auerbach and Feldstein (1987).

Olson, M. 1965. *The logic of collective action*. Cambridge, MA: Harvard University Press.

Pigou, A.C. 1928. *A study in public finance*. 3rd ed. London: Macmillan.

Rubinfeld, D.L. 1987. The economics of the local public sector. In Auerbach and Feldstein (1987).

Samuelson, P.A. 1954. The pure theory of public expenditure. *Review of Economics and Statistics* 36: 387–389.

Samuelson, P.A. 1955. Diagrammatic exposition of a theory of public expenditure. *Review of Economics and Statistics* 37: 350–356.

Samuelson, P.A. 1969. Pure theory of public expenditure and taxation. In *Public economics*, ed. J. Margolis and H. Guitton. London: Macmillan.

Sandmo, A. 1972. Optimality rules for the provision of collective factors of production. *Journal of Public Economics* 1: 149–157.

Sandmo, A. 1973. Public goods and the technology of consumption. *Review of Economic Studies* 40: 517–528.

Sandmo, A. 1998. Redistribution and the marginal cost of public funds. *Journal of Public Economics* 70: 365–382.

Sandmo, A. 2003. International aspects of public goods provision. In *Providing global public goods*, ed. I. Kaul et al. New York: Oxford University Press.

Sax, E. 1924. The valuation theory of taxation. In Musgrave and Peacock (1958).

Scotchmer, S. 2002. Local public goods and clubs. In *Handbook of public economics*, ed. A.J. Auerbach and M. Feldstein, vol. 4. Amsterdam: North-Holland.

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

Tiebout, C.M. 1956. A pure theory of local expenditures. *Journal of Political Economy* 64: 416–424.

Tulkens, H. 1978. Dynamic processes for public goods: An institution-oriented survey. *Journal of Public Economics* 9: 163–201.

Vickrey, W.S. 1961. Counterspeculation, auctions, and competitive sealed tenders. *Journal of Finance* 16: 8–37. Repr. in W.S. Vickrey, *Public economics*, ed. R. Arnott, K. Arrow, A. Atkinson and J. Drèze. Cambridge: Cambridge University Press, 1994.

Warr, P. 1983. The private provision of a public good is independent of the distribution of income. *Economics Letters* 13: 207–211.

Wicksell, K. 1896. *Finanztheoretische Untersuchungen*. Partially translated as 'A new principle of just taxation'. In Musgrave and Peacock (1958).

Public Goods Experiments

Rachel T. A. Croson

Abstract

Economic theory often cites the existence of goods with externalities as justification for government intervention, either as taxation to fund goods with positive externalities which would otherwise be underprovided, or as regulation on goods with negative externalities which would otherwise be overprovided. A series of experiments tests these predictions of under- or over-provision. This article describes the landscape of public goods experiments, identifying similarities and differences between them and summarizing the broad findings.

Keywords

Altruism; Common pool resource game; Cooperation; Coordination; Externalities; Learning; Linear public goods game; Other-regarding preferences; Prisoner's Dilemma; Production functions; Provision point mechanism; Public good experiments; Public goods; Reciprocity; Rent seeking; Repeated games; Self-interest; Social cost; Tragedy of the commons; Voluntary contribution mechanism

JEL Classifications

C9

Since the vivid description of the Prisoner's Dilemma game and the accompanying tension between self-interest and efficiency, economists, political scientists, psychologists, sociologists and others have wondered about how individuals resolve these conflicting motivations. Many have investigated this question by experimentally examining behaviour when individual interest and group interest conflict. This research goes under different names – for e7xample, social dilemmas in psychology, commons dilemmas in political science and public goods problems in economics.

This article does not provide a comprehensive review of this research; other excellent reviews aimed at economists (Ledyard 1995), psychologists (Dawes 1980) or sociologists (Kollock 1998) exist. Instead, I highlight the different categories of public goods problems and introduce their commonalities.

The rest of this article is organized as follows. First, I offer some definitions of characteristics that public goods problems share. Next I discuss some dimensions on which they differ, and how these dimensions translate into different equilibrium and efficient outcomes. Then I describe three specific public goods problem types that have been extensively studied in the economics literature: the voluntary contribution mechanism, the provision point mechanism and the common pool resource. I conclude with a description of other games that have been studied, but where more work could be done.

Similarities

The common feature in public goods settings is the existence of externalities. In public GOODS problems, individuals can use private resources to provide goods that have positive externalities for others. Since some social benefits are not captured by the individual making the decision, this results in under-provision relative to the socially optimum level. Self-interested economic theory, then, argues that these goods will be under-provided, justifying taxation as a role for government.

Parallel to the public goods situation is the case of public BADS. Here, individuals receive private resources by producing goods that have a negative externality for others. Since some social costs are not captured by the individual making the decision, this results in overprovision of public bads relative to the socially optimum level. Self-interested economic theory, then, argues that these goods will be over-provided, again justifying a role for government, here in regulation. (A number of models have extended the existing theory to internalize the externalities. In models of altruism – for example, Becker 1974; Andreoni 1989 – the utility function of one party includes the consumption of others. Thus, when an action creates positive externalities, the value from it is increased over the self-interested model. Charness and Rabin 2002, posit that individuals care not only about their own consumption but also about social welfare directly. These other- regarding preferences can internalize some of the externalities in public goods problems, but typically the over- or under-provision problems are not eliminated.)

The defining characteristic of the public goods problems described here, however, is the existence of positive externalities. My actions affect others, and I do not take this effect (sufficiently) into account in my own maximization problem. Often these problems are symmetric; each individual faces the identical conflict, but this is not necessary for there to be a public goods problem. For a problem to exist, it must only be that the individual's welfare and the group's (social) welfare conflict.

Differences

The largest, most important and least-recognized difference in public goods situations is the production function. How does an individual's action create positive or negative externalities for others? Different production functions have different implications for equilibrium predictions as well as socially optimal outcomes.

A second important dimension on which these situations differ is the decision space. Public goods problems can involve acting to provide goods with positive externalities at a private cost, refraining from acting so as to avoid imposing negative externalities on others at a personal cost, or acting to capture what would otherwise be public benefits for private consumption. Unlike the first dimension, these differences in the decision space have only a superficial impact on the equilibria; that is, one can easily describe 'not polluting' as 'producing a public good'. However, they may affect how individuals think about (and act in) these problems.

To clarify these differences, let's examine the three classic games discussed below in terms of the production function and decision space. In the voluntary contribution mechanism (VCM), the decision space involves *GIVING.* Individuals are given an endowment, which they can use for their private consumption or to produce the public good. Their allocations toward the public good provide value for others in the experiment (positive externalities). The production function of the VCM game most extensively studied is *LINEAR* (thus it is sometimes also called a linear public goods game). The more that is allocated toward the public good, the greater are the social benefits in a linear fashion. This linearity means that (with appropriate parameters as discussed below) this game has a unique Nash equilibrium in which no participant allocates any resources towards producing the public good. Deviations from the equilibrium are both welfare-enhancing and represent deviations from pure self-interest maximization. They are thus referred to as 'cooperation', and concepts like altruism, warm glow and reciprocity are offered as their explanation.

In contrast, consider the provision point mechanism (PPM). Again, this game typically has a *GIVING* decision space. Individuals are given an endowment which can be allocated to private consumption or towards providing the public good. But, in contrast with the VCM, in the PPM the production function involves a *THRESHOLD.* If enough resources are collected, then the public good is provided and all receive its benefits. If too few resources are collected, then the public good is not provided and no positive externalities are enjoyed. The threshold nature of this production function has critical implications for the equilibria of this game. With appropriate parameters (discussed below), the full free-riding equilibrium still exists. But there also exist a set of efficient equilibria, in which the public good is exactly provided with each individual contributing a share of its cost. In each of these equilibria, the share contributed by each individual varies. (For example, there may be one equilibrium in which I contribute 80 per cent and you contribute 20 per cent of the cost of providing the public good, and another where I contribute 20 per cent and you contribute 80 per cent.) The problem then becomes one not of cooperation but primarily one of coordination; how do we select among these efficient equilibria? Formally, this game can be thought of as a large battle-of-the-sexes game (a game of impure coordination), with multiple equilibria each of which is somebody's favourite.

Finally, consider the common pool resource (CPR) game. Here the decision space involves *TAKING*; for example, individuals can harvest grass from the commons for personal gain. Surprisingly, these experiments are often described as *GIVING* games, with negative externalities rather than positive, as in the VCM or PPM. So the decision is made on 'how many hours to spend grazing' with the resulting negative externalities as cattle eat more grass. The production function used in CPR games is typically *NONLINEAR.* A small amount of harvesting creates more benefit for the individual than harm to the society (and is thus socially efficient). However, as the individual harvests more, the personal benefits decrease and the social costs increase until societal costs outweigh private benefits (the socially optimal point). With appropriate parameters (described below), however, private benefit is still above private costs, leading individuals to continue harvesting past the socially efficient point. Eventually, private benefits equal private costs, leading individuals to stop harvesting. These equilibria are thus internal (individuals typically harvest more than zero but less than the full amount) but still suboptimal (total harvesting is larger than the socially optimum level).

There are many additional dimensions on which these games can vary. For example, experimenters have varied the number of players from two (the classic Prisoner's Dilemma game) to as many as 100 (Isaac et al. 1994). The particular parameters can vary, subject to constraints that preserve the public goods nature of the problem. The institutional rules can vary, participants can decide simultaneously or sequentially, they can discuss the game in advance or not, and so on. The games can be (finitely) repeated or one-shot. I discuss some of these variations in the sections below, but their impacts on the equilibria of the games are straightforward.

In summary, the set of public goods games is broad. When one looks at a game, however, it is critical to understand the production function that is being used to translate decisions into outcomes (positive/negative, linear/threshold/nonlinear), and the decision space that participants face (giving/taking/refraining from action). These dimensions have important impacts on the equilibrium predictions, the observed behaviour and the attributions that one can make about the causes of differences between the two.

The Voluntary Contribution Mechanism (VCM)

The work of Marwell and Ames (1979, 1980, 1981) is often cited as the earliest VCM experiments. Unfortunately these early experiments did not involve a linear production function. Instead the return from the public account was discrete (chunky) although some of the experiments involved a linear approximation (for example, 1981, study I). Furthermore, the experiments were relatively uncontrolled; subjects had instructions mailed to them at home, were individually called and had the instructions explained to them, and then called back one week later and made their (one-shot) decision by phone.

The first paper using a linear VCM in a controlled lab setting was Isaac et al. (1984). This paper set a number of precedents for how such experiments are run. In this experiment, participants were brought into the lab and arranged into fixed groups of four. In each period, each group member was given tokens, which he could allocate between a private account and a group account. Tokens allocated to the private account earned 1¢ per token. Tokens allocated to the group account earned 0.3¢ per token for each member of the group, whether or not he had contributed to the group account. As the production function is linear, these parameters remain constant regardless of how much is contributed.

More generally, for there to be a public goods problem in these linear games, a few conditions must be satisfied. First, the return from the public good to the individual must be lower than the return from the private good ($0.3 < 1$). This ensures that individuals do not have an individual incentive to contribute, and that the dominant strategy equilibrium in the stage game is thus to contribute zero tokens. Furthermore, the social benefit from contributing toward the public good must be greater than the social cost ($0.3*4 = 1.2 > 1$). This ensures that contributing toward the public good is socially efficient.

The game is finitely repeated for ten periods, to allow for convergence to (and learning of) the equilibrium. In the finitely repeated game, backward induction results in the unique Nash equilibrium of zero contributions. Deviations from that equilibrium are attributed to cooperation, altruism, reciprocity or various other- regarding preferences.

A number of precedents set in this original article have been used in subsequent research.

Many papers use a group size of four, although some have gone as low as two and others as high as 100. Most experiments have participants 'allocate' tokens between multiple funds rather than 'contribute' towards a public good, as this experiment did. Participants typically have multiple tokens to allocate rather than simply one. Most papers use repetition with fixed groups, and many choose ten periods.

The results from this wide variety of experiments are quite robust. First, on average, contributions to the public good begin at about half the endowment of tokens. Second, there is considerable variation in the decisions of individuals. Third, those contributions reduce over time until the contributions in the final round are 10–20 per cent of the endowment. An example of this pattern of contributions is depicted in Fig. 1. A number of interesting papers have hypothesized and tested for the source of these regularities. Some explanations include errors (Palfrey and Prisbrey 1997), confusion (Andreoni 1995b), strategies and learning (Andreoni and Croson 2008), and reciprocity or conditional cooperation (Croson 2007), among others.

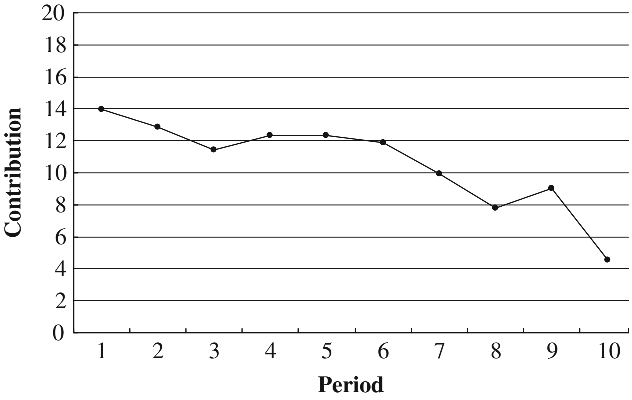
Variations in the parameters have been explored as well; individual papers manipulate group size (Isaac and Walker 1988b), the ratio of the return from the public good to the return from the private good (Isaac and Walker 1988b), the existence of communication (Isaac and Walker 1988a), fixed groups (Andreoni and Croson 2008), anonymity (Laury et al. 1995) and framing (Andreoni 1995a). Recent work in this area extends the paradigm to incorporate more realistic assumptions, including heterogeneity of players (Buckley and Croson 2006), endogenous group formation (Croson et al. 2005), and punishment/reward (Fehr and Gächter 2000). Data has been collected from various subject pools, including children (Krause and Harbaugh 2000) and residents of Asian slums (Carpenter et al. 2004). (For a fascinating look into underappreciated but related psychology literature, see research on *SOCIAL LOAFING*, reviewed in Karau and Williams 1993.)

In summary, the VCM captures the pure tension between individual gains and social efficiency. It is thus used in many settings and by many researchers to investigate the causes (and consequences) of this tension, as well as to describe behaviour in the world.

The Provision Point Mechanism (PPM)

One concern with the VCM is that in equilibrium the public good is not provided at the socially efficient level. Bagnoli and Lipman (1989) discuss a logical response to this problem: add a threshold (or provision point) to the production process. The threshold needed to provide the public good is announced. If at least that much is allocated to the group account, then the public good is produced; if not, no public good is produced.

It is straightforward to see that a VCM can be 'discretized' to the PPM by adding a threshold.



Public Goods Experiments, Fig. 1 Average contributions to public account in VCM. *Source:* Croson (2007)

With the appropriate parameters (discussed below) this game now has a set of efficient Nash equilibria in which the public good is exactly provided. There are also inefficient equilibria of this game, in which the public good is not provided, but this mechanism nonetheless represents a theoretical improvement over the VCM.

There are some parameter values necessary for the existence of these efficient equilibria. In particular, imagine the threshold is T, the value from private consumption is 1 and individual endowments are $E_{\rm i}$. Define V_I as an individual's value from the public good. For an efficient equilibrium there must exist a set of contributions $\{\sigma_i\}$ such that $\Sigma\sigma_i \geq T$. Furthermore, the individual rationality constraints must be satisfied $(\forall_I)$ $\sigma_i \leq \min\{E_i,v_i\}$ and providing the public good must be efficient $T \leq \Sigma v_i$.

Additional assumptions are needed before this mechanism is completely described. When the threshold is not reached, the resources contributed to it can be returned or can be lost. This feature has been called the 'money back guarantee' in psychology, and in economics is the *refund* (Isaac et al. 1989; Bagnoli and McKee 1991). The existence of a refund does not affect the set of efficient equilibria, but does change the set of inefficient equilibria. With no refund, there is one (unique) inefficient equilibrium of zero contribution. With a refund, there are many (weak) inefficient equilibria in which some is contributed towards the public good, but not so much that any player can supplement to reach the threshold. Those contributions are then refunded, making the contributors indifferent between these strategies and contributing zero.

The second dimension is the disposition of resources above the threshold. This is referred to as the *rebate* (Marks and Croson 1998). Experiments have been run including no rebate (excess contributions are lost), proportional rebates (excess contributions are returned proportionally based on contributions), and utilization rebates (excess contributions are used to provide the public good in a VCM fashion). None of these changes the set of equilibria.

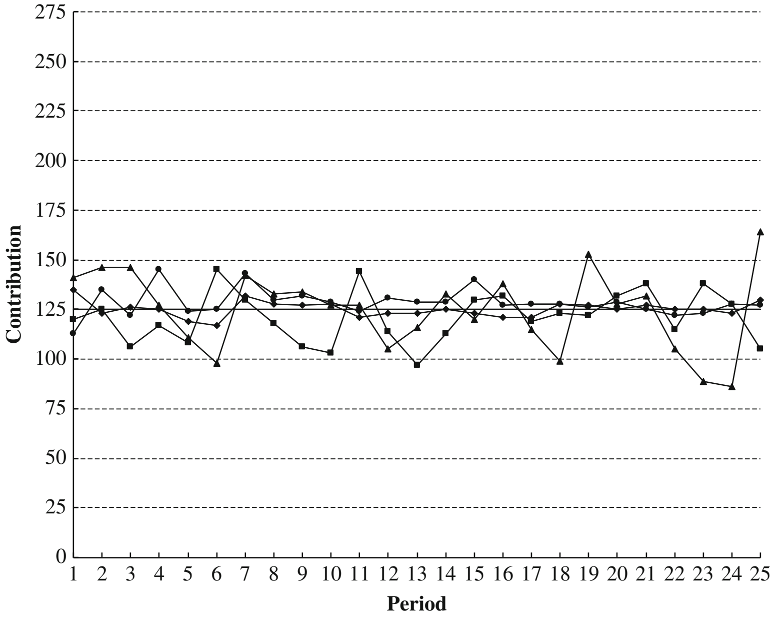
While the PPM has the advantage of the existence of efficient equilibria, it has the disadvantage of too many equilibria. For example, in a typical parameterization used by Croson and Marks (1998), five players each had 55 tokens to allocate. Tokens allocated to the private account earned 1¢ each. If there were at least 125 tokens allocated to the public account, each participant in the group received 50¢. These parameters satisfy the conditions above; the collective benefit from the public good (5 people × 50¢ = \$2.50) is greater than the social cost of provision (\$1.25). There exists a set of allocations such that the public good is provided; one is the unique symmetric equilibrium in which each player allocates 25 tokens, the threshold is exactly met, and each participant receives their value of 50¢, strictly greater than their costs of 25¢.

Unfortunately, this is not the only efficient equilibrium. In particular, the set of allocations {25, 25, 25, 24, 26} is also an equilibrium, as is {25, 25, 25, 26, 24}, although player 4 prefers the former and player 5 the latter. All told, there are 4,052,751 efficient equilibria using these parameters. Thus the main problem in the PPM is not one of *cooperation;* avoiding the inefficient outcome as in the VCM. It's a problem of *coordination*, of choosing which of the many efficient equilibria the group will play.

The coordination problem is difficult enough in the stage game. However, in the lab this game is typically finitely repeated. In the repeated game, the number of potential equilibria grows exponentially, as any sequence of stage-game equilibria are themselves an equilibrium of the repeated game.

In practice, almost no instances of the inefficient equilibria are observed. Group contributions tend to cycle around the efficient equilibrium level, although they are almost equally likely to be above the threshold as below. Examples of group contributions in the PPM can be seen in Fig. 2.

Further research has investigated other dimensions of the PPM. These include the effect of subject pool (Cadsby and Maynes 1998), binary versus continuous giving (Cadsby and Maynes 1999), heterogeneous valuations (Croson and Marks 1999), identifiability of contributions (Croson and Marks 1998), incomplete



Public Goods Experiments, Fig. 2 Group contributions to public account in PPM, four groups. *Source:* Croson and Marks (2000)

information (Marks and Croson 1999), and framing (Sonnemans et al. 1998).

The PPM has a number of important and interesting properties. It allows for efficient equilibria, thus to some extent 'solving' the public goods problem. However, this solution brings costs: too many equilibria and the need to coordinate among them. This distinction, between the cooperation motive of the VCM and the coordination motive of the PPM is a critical and often-overlooked one.

The Common Pool Resource Game

The structure of the CPR game is based on work by Gordon (1954) and Hardin (1968) on the tragedy of the commons. In the typical tragedy, ranchers graze their herds either on their private land or on the commonly owned land in each village. Since grazing on the commons is free, individuals prefer it to using their own land, which can be used to grow cash crops. However, grazing imposes a negative externality on others; if my cows eat the grass, there is less left for your herd. The CPR game, thus, is a *CONTINUOUS TAKING* game; each unit of grass that I take exerts a negative externality on the rest of the village.

Unlike the VCM, the externalities imposed are typically nonlinear, with public costs initially being lower than private benefit, but rising until the two cross. Thus the game has internal equilibria, in which more grazing than is optimal is predicted. (These games are similar to a class of *RENT-SEEKING* games, which have recently been experimentally explored. Rent-seeking games are beyond the scope of this article; but see Önçüler and Croson 2005, for some recent work.)

In the first CPR economics experiment, Walker et al. (1990) arranged subjects into groups of eight. Each participant was given a homogeneous endowment and was told he could allocate this endowment between two markets. Like the VCM, the private market paid a fixed amount, 5b per token. The public market (the common pool) had externalities for other group members' consumption. Unlike the VCM this externality was negative rather than positive. Also unlike the VCM, the externality was nonlinear, with increasing social cost. Conceptually, allocating resources to the public market captures the idea of grazing the herd on public land.

When x_i is the amount player *I* allocates to the public market, the earnings from the public market for player *I* are:

$$x_i / \sum x_i \left(23 \sum x_i - 0.25 \left(\sum x_i \right)^2 \right).$$

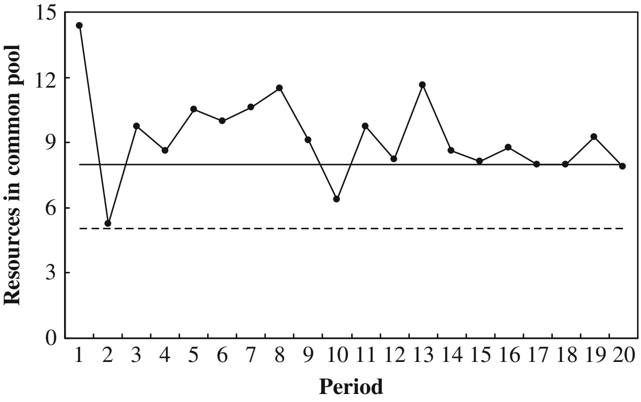
The negative squared term creates the nonlinearity. If no one is allocating resources to the public market, an individual earns more from that market than the private one (the first token allocated there earns 22.5¢ versus 5¢ in the private market). However, this return quickly diminishes, so the value from investing in the public market falls below the value from investing in the private market as the number of tokens increases. This captures the negative externalities. For each token that player *i* invests in the public market, the marginal value of player's *j's* investment in that market is lowered.

The self-interested, symmetric Nash equilibrium in this game is for each player to invest eight tokens in the public market (for a total investment of 64 tokens). (When each participant invests nine tokens in the public market, the return for that marginal token is exactly 5¢. The authors assume that, when indifferent participants choose not to impose negative externalities on others, thus the equilibrium of eight tokens is used.) This equilibrium prediction is parallel to the prediction of full free-riding in the VCM. In contrast, the symmetric, socially efficient solution is for each participant to invest five tokens in the public market. This is not an equilibrium, however, since each individual privately captures more by investing further in the public market. This capturing is at the expense of the other players, who suffer the negative externality imposed. So five tokens is the socially optimal level, and is parallel to the prediction of full contributing in the VCM.

If behaviour in the CPR were parallel to that in the VCM, we should see allocations to the common market of between eight tokens (the equilibrium) and five tokens (the social optimum). As in the VCM, the stage game described above is repeated finitely many times, either 20 or 30 rounds, depending on the particular parameters. (A parallel literature in CPR games examines *dynamic* versions of the game, in which the resource replenishes itself round to round, with the replenishment rate being dependent on the harvesting rate observed. These are sometimes referred to as *renewable* CPR games. Equilibria in these dynamic games are more complicated, and Herr et al. 1997, experimentally compare the different games.)

The results from the experiment can be seen in Fig. 3. The solid line represents the equilibrium prediction, while the dotted line represents the social welfare maximizing outcome. Unlike the VCM, where contributions lay between these two, here contributions lie on the opposite side of the equilibrium. This indicates excessive allocation to the public market, and excessive negative externalities, over and above the equilibrium prediction.

This result of less-than-Nash levels of cooperation is replicated in other experiments, reviewed in Ostrom et al. (1994). Other work also reviewed there examines other questions in CPR games, including probabilistic destruction, communication, monitoring and sanctions, voting and heterogeneity.



Public Goods Experiments, Fig. 3 Average resources allocated to the common pool (CPR). *Source:* Ostrom et al. (1994)

One lingering puzzle remains: why are subjects more generous/cooperative than the equilibrium in the VCM and less generous/cooperative than the equilibrium in the CPR game? A number of studies have investigated this question by adding complexity to the VCM to make it resemble the CPR (for example, the stream of research on nonlinear VCM games below). Others investigate framing, suggesting it is the difference between providing a positive externality in the VCM and a negative externality in the CPR game. Unfortunately no study has offered either a definitive experiment or compelling data to explain why the outcomes from these games differ.

Other Public Goods Settings

In addition to the games described above, a small literature explores different types of public goods games. A number of papers examine nonlinear VCMs, with internal equilibria (see Laury and Holt 2008, for a review). Here the production of public good is nonlinearly related to the amount allocated to the public account. This yields an internal social optimum and Nash equilibrium level of contributions. As before, parameters are set so that the public good is under-provided in equilibrium.

Others have explored markets with externalities rather than public goods per se (for example, Plott 1983). Still other researchers combine these games in creative ways, for example a PPM with a VCM for excess contributions (as in the utilization rebate of Marks and Croson 1998), or a PPM with a VCM for under-contributions (as in Vesterlund et al. 2005).

Finally, a number of papers have experimentally tested other proposed mechanisms for solving the public goods problem. For example, Chen and Plott (1996) provide a test of the Groves–Ledyard mechanism (a mechanism designed to elicit individuals values for public goods). Reviews of experiments using incentive-compatible mechanisms can be found in Chen (2008). These literatures are less developed than the previous three games, a disadvantage when trying to summarize a stream of research but an advantage when seeking a new contribution.

Commonalities and Puzzles

The underlying similarity between all public goods experiments is the existence of externalities. These externalities can be positive or negative, and they can be linear, nonlinear or involve thresholds. The decisions participants make can be described as giving or taking. These varying situations affect the equilibrium predictions of the games.

Individuals are ‘cooperative’ in the VCM; they contribute more towards the public good than equilibrium behaviour would predict. There are many explanations for why this may be the case, including altruism, reciprocity (conditional altruism), warm-glow and errors, but no one causal factor has emerged as dominant.

In the PPM, the issue is not one of cooperation but coordination. On average the efficient equilibrium outcomes describe the data. However, there is also ‘gaming’, with groups sometimes failing to provide the public good as one individual attempts to move towards a more attractive equilibrium. Thus, while outcomes from these mechanisms are more efficient than those from the VCM, the coordination problem is severe and unsolved.

Finally, individuals harvest *MORE* than the Nash equilibrium predictions in CPR games. This result contrasts with the VCM; here individuals are more competitive than the equilibrium prediction. The source of these differences is still unexplored and represents an excellent direction for future research.

Summary

The tension between self-interest and social efficiency is one we experience every day. Experiments like those discussed in this article have been developed to explore how humans resolve this tension. Results from these experiments highlight the impact of different public goods structures, institutional arrangements and repeated interactions on human behaviour. Ultimately they help us to design mechanisms to better provide public goods, and allow for a deeper understanding of human motivations in the wide set of activities involving externalities for others.

See Also

▶ Altruism in Experiments
▶ Common Property Resources
▶ Coordination Problems and Communication
▶ Experimental Economics
▶ Public Goods
▶ Reciprocity and Collective Action

Bibliography

Andreoni, J. 1989. Giving with impure altruism: Applications to charity and Ricardian equivalence. *Journal of Political Economy* 97: 1447–1458.

Andreoni, J. 1995a. Warm-glow versus cold-prickle: The effects of positive and negative framing on cooperation in experiments. *Quarterly Journal of Economics* 110: 1–21.

Andreoni, J. 1995b. Cooperation in public-goods experiments: Kindness or confusion? *American Economic Review* 85: 891–904.

Andreoni, J., and R.T.A. Croson. 2008. Partners versus strangers: Random rematching in public goods experiments. In Plott and Smith (2008).

Bagnoli, M., and B.L. Lipman. 1989. Provision of public goods: Fully implementing the core through private contributions. *Review of Economic Studies* 56: 583–601.

Bagnoli, M., and M. McKee. 1991. Voluntary contribution games: Efficient provision of public goods. *Economic Inquiry* 29: 351–366.

Becker, G.S. 1974. A theory of social interactions. *Journal of Political Economy* 82: 1063–1093.

Buckley, E., and R.T.A. Croson. 2006. Income and wealth heterogeneity in the voluntary provision of linear public goods. *Journal of Public Economics* 90: 935–955.

Cadsby, B., and E. Maynes. 1998. Choosing between a socially efficient and free-riding equilibrium: Nurses versus economics and business students. *Journal of Economic Behavior and Organization* 37: 183–192.

Cadsby, B., and E. Maynes. 1999. Voluntary contribution of threshold public goods with continuous provisions: Experimental evidence. *Journal of Public Economics* 71: 53–73.

Carpenter, J.P., A.G. Daniere, and L.M. Takahashi. 2004. Cooperation, trust, and social capital in Southeast Asian urban slums. *Journal of Economic Behavior and Organization* 55: 533–551.

Charness, G., and M. Rabin. 2002. Understanding social preferences with simple tests. *Quarterly Journal of Economics* 117: 817–869.

Chen, Y. 2008. Incentive-compatible mechanisms for pure public goods: A survey of experimental literature. In Plott and Smith (2008).

Chen, Y., and C.R. Plott. 1996. The Groves–Ledyard mechanism: An experimental study of institutional design. *Journal of Public Economics* 59: 335–364.

Croson, R.T.A. 2000. Feedback in voluntary contribution mechanisms: An experiment in team production. *Research in Experimental Economics* 8: 85–97.

Croson, R.T.A. 2007. Theories of commitment, altruism and reciprocity: Evidence from linear public goods games. *Economic Inquiry* (forthcoming).

Croson, R.T.A., and M. Marks. 1998. Identifiability of individual contributions in a threshold public goods experiment. *Journal of Mathematical Psychology* 42: 167–190.

Croson, R.T.A., and M. Marks. 1999. The effect of heterogeneous valuations for threshold public goods: An experimental study. *Risk, Decision and Policy* 4: 99–115.

Croson, R.T.A., and M. Marks. 2000. Step returns in threshold public goods: A meta- and experimental analysis. *Experimental Economics* 2: 239–259.

Croson, R.T.A., and M. Marks. 2001. The effect of recommended contributions in the voluntary provision of public goods. *Economic Inquiry* 39: 238–249.

Croson, R.T.A., E. Fatas, and T. Neugebauer. 2005. Excludability in three public goods games. Working Paper, Wharton School, University of Pennsylvania.

Dawes, R. 1980. Social dilemmas. *Annual Review of Psychology* 31: 169–193.

Fehr, E., and S. Gächter. 2000. Cooperation and punishment in public goods experiments. *American Economic Review* 90: 980–994.

Gordon, H.A. 1954. The economic theory of a common property resource: The fishery. *Journal of Political Economy* 62: 124–142.

Hardin, G. 1968. The tragedy of the commons. *Science* 162: 1243–1248.

Herr, A., R. Gardner, and J.M. Walker. 1997. An experimental study of time-independent and time-dependent externalities in the commons. *Games and Economic Behavior* 19: 77–96.

Isaac, R.M., and J.M. Walker. 1988a. Communication and free-riding behavior: The voluntary contribution mechanism. *Economic Inquiry* 26: 585–608.

Isaac, M.R., and J.M. Walker. 1988b. Group size effects in public goods provision: The voluntary contributions mechanism. *Quarterly Journal of Economics* 53: 179–200.

Isaac, R.M., J.M. Walker, and S.H. Thomas. 1984. Divergent evidence on free riding: An experimental examination of possible explanations. *Public Choice* 43: 113–149.

Isaac, R.M., D. Schmidtz, and J.M. Walker. 1989. The assurance problem in a laboratory market. *Public Choice* 62: 217–236.

Isaac, R.M., J.M. Walker, and A.W. Williams. 1994. Group size and the voluntary provision of public goods: Experimental evidence utilizing large groups. *Journal of Public Economics* 54: 1–36.

Karau, S.J., and K.D. Williams. 1993. Social loafing: A meta-analytic review and theoretical integration. *Journal of Personality and Social Psychology* 65: 681–706.

Kollock, P. 1998. Social dilemmas: The anatomy of cooperation. *Annual Review of Sociology* 24: 183–214.

Krause, K., and W.T. Harbaugh. 2000. Children's contributions in public good experiments: The development of altruistic and free-riding behaviors. *Economic Inquiry* 38: 95–109.

Laury, S.K., and C.A. Holt. 2008. Voluntary provision of public goods: Experimental results with interior Nash equilibria. In Plott and Smith (2008).

Laury, S.K., J.M. Walker, and A.W. Williams. 1995. Anonymity and the voluntary provision of public goods. *Journal of Economic Behavior & Organization* 27: 365–380.

Ledyard, J. 1995. Public goods: A survey of experimental research. In *Handbook of experimental economics*, ed. J. Kagel and A. Roth. Princeton: Princeton University Press.

Marks, M., and R.T.A. Croson. 1998. The effect of alternative rebate rules in the provision point mechanism of voluntary contributions: An experimental investigation. *Journal of Public Economics* 67: 195–220.

Marks, M., and R.T.A. Croson. 1999. The effect of incomplete information and heterogeneity in the provision point mechanism of voluntary contributions: An experimental investigation. *Public Choice* 99: 103–118.

Marwell, G., and R.E. Ames. 1979. Experiments on the provision of public good I: Resources, interest, group size, and the free-rider problem. *American Journal of Sociology* 84: 1336–1360.

Marwell, G., and R.E. Ames. 1980. Experiments on the provision of public goods II: Provision points, stakes, experience, and the free-rider problem. *American Journal of Sociology* 85: 926–937.

Marwell, G., and R.E. Ames. 1981. Economists free ride, does anyone else? Experiments on the provision of public goods, IV. *Journal of Public Economics* 15: 295–310.

Önçüler, A., and R.T.A. Croson. 2005. Rent-seeking for a risky rent: A model and experimental investigation. *Journal of Theoretical Politics* 17: 403–429.

Ostrom, E., R. Gardner, and J.M. Walker. 1994. *Rules, games, and common-pool resources*. Ann Arbor: University of Michigan Press.

Palfrey, T.R., and J.E. Prisbrey. 1997. Anomalous behavior in public goods experiments: How much and why? *American Economic Review* 87: 829–846.

Plott, C.R. 1983. Externalities and corrective policies in experimental markets. *Economic Journal* 93: 106–127.

Plott, C.R., and V.L. Smith, eds. 2008. *Handbook of experimental economics results.* Amsterdam: North-Holland (forthcoming).

Sonnemans, J., A. Schram, and T. Offerman. 1998. Public good provision and public bad prevention: The effect of framing. *Journal of Economic Behavior & Organization* 34: 143–161.

Vesterlund, L., J. Duffy, and J. Ochs. 2005. Giving little by little: Dynamic voluntary contribution games. Working Paper, University of Pittsburgh.

Walker, J.M., R. Gardner, and E. Ostrom. 1990. Rent dissipation in a limited-access common-pool resource: Experimental evidence. *Journal of Environmental Economics and Management* 19: 203–211.

Public Health

Alan Williams

It is widely believed that prevention is better than cure, and there is an obvious sense in which it is. But there has been a chastening tendency for the dismal science to indicate that measures enthusiastically advocated as ways of stopping people from becoming ill are often rather poor investments compared with 'curative' activities.

What Is 'Public Health'?

Before getting to grips with that particular issue, however, we need to give some attention to what constitutes 'public health' as a distinctive activity. Several characteristics seem relevant. Public health could be concerned with the health of whole populations rather than with that of individuals, in which case 'public' is being contrasted with 'individual'. Or it could be concerned with measures that *have* to be applied to whole communities (e.g. fluoridation of water supplies) rather than to individuals (e.g. immunization) hence 'public' is being contrasted with 'personal'. Or it could be concerned with all *preventative* activity, as opposed to *curative* activity, so that the contrast is rather between a positive emhasis on promoting public *health* in contrast with alleviating public *illness*. Some 'public health' activities (e.g. ensuring that water supplies are wholesome and sewage disposal facilities are functioning effectively) clearly embody all three attributes, but what about anti-smoking campaigns? For the purpose of this particular discussion, public health will be held to embrace *any* governmental activity designed to keep healthy people healthy. This would include counselling people about their lifestyle (smoking, alcohol, diet, exercise, etc.), plus all safety legislation, and much consumer protection, as well as such 'traditional' areas as protection from environmental pollution, immunization and vaccination programmes, and the control of epidemics. They are rather disparate activities, and many of them do not lie within the field of responsibility of those providing personal health services, nor indeed of Ministries of Health, so that organizationally they manifest a rather complex picture of overlapping responsibilities in some fields, and rather yawning gaps in others, so their organizational aspects may be of as much interest as the analytical problems they pose.

What Is Wrong with the Market?

Since it is impossible to cover this whole complex territory in any depth in a short essay such as this, a few representative problems will be selected to give the flavour of the *economic* issues involved. These economic issues can most conveniently be organized around the notion of 'market failure', though this in turn leads inexorably to notions of 'government failure' too, if we are to pick up the organizational difficulties.

Generally we economists have a bias in favour of market solutions, once we are satisfied that consumers are fairly knowledgeable (and the consequences of ignorance are not likely to be fatal), that they are properly regarded as the best judges of their own interests, that when acting in their own interests they are not likely to affect other people's interests in unacceptable ways, that consumers are not subject to exploitation by suppliers (or vice versa) through disproportionate market power, and that the distribution of power (and the distribution of welfare which flows from it) is equitable. These are quite stringent conditions, and should not be taken for granted (i.e. there should be no unscrutinized *presumption* that market solutions are best).

Consumer Ignorance

Consumer ignorance is thus a possible basis for public health activity directed at changing people's diet, promoting physical exercise, and launching anti-smoking campaigns. It also lies behind much consumer protection activity, such as that concerned to prevent adulteration of food and drink, and with safety of other products from the viewpoint of fire-risk, the use of toxic materials, electrical safety, etc. Although there are *private* consumer protection associations offering such advisory services, they suffer from the weakness that only those who are already aware of such risks in a general way will be prepared to pay for such advice, and once products have been shown to be health hazards the ethos of such private organizations makes them wish to disseminate such information as widely as possible, so there is an obvious 'free-rider' problem. Moreover, powerful producer interests may unscrupulously protect themselves by blocking adverse publicity in the media by using the market power of advertising revenues. Moreover, through superficial product differentation, rapid product 'development' may enable may adverse effects of bad reports to be sloughed off, and it therefore requires continuous, prompt, and expensive reappraisal of products to keep consumers well informed. It is, therefore, not surprising that the coercive power of the government is sought to prevent these dangers from presenting themselves in the first place, by banning certain food additives, setting minimum safety standards for products, and enforcing these by law. Problems of inspection and testing then fall in the public domain, and their rigour and effectiveness depends on the resources devoted to them and the attitudes of the inspectorate and of the courts on which they ultimately depend. There is no guarantee that such regulations will be 'optimal' in the sense that the social costs and benefits at the margin will be roughly equal. For instance, it appears that fire and construction safety regulations for buildings imply much higher values being placed on human life than do road safety measures, so there is a further field for economic analysis here to rationalize piecemeal and often quite impulsive legislation (often sparked off by some disaster or scandal).

Externalities

The next feature which often seems to underlie public health activity is the 'externality' argument, i.e. that individuals, behaving rationally in their own interests, will ignore the interests of others, and it may be in the interests of all to coerce each and every one to behave in a different manner. On the external cost side this has been the classic rationale for the isolation of suffers from infectious and contagious diseases. But it has some more subtle and pervasive applications, e.g. to drunken driving, or to smoking in enclosed public places or to failure to maintain motor vehicles in a roadworthy condition, where individuals may rationally choose to accept risks themselves, but in so doing put others at risk involuntarily and without offering proper compensation for accepting such risks. This then raises the issue of how far we are able to go in using *coercive* measures to control people's behaviour for the sake of other people's health. Should people be compelled to wash thoroughly everyday, or not wear dirty clothing on public transport, since one person's low standard of cleanliness may endanger someone else's health? Extending the point still further, have I the right to insist that an ill person seeks treatment for an illness which I cannot catch, simply because it causes me distress to see that person ill when I know he or she could be made well? The externality argument, if pushed hard, is extremely intrusive, and finding an acceptable balance between protecting ourselves from other people's irresponsibility and protecting our own private area of responsibility from outside interference is consequently a much debated issue in this as in other fields of human activity.

The externality argument applies equally strongly in principle on the benefit side, of course, and the classic case is the draining of the malarial swamp. Most such instances are closer to 'local' public goods (i.e. ones that have a strong spatial dimension) than to 'pure' public goods, but they do have the key characteristic that the benefit is equally available to all (within some designated geographical area) and exclusion from benefit (conditional upon paying for it) is impossible. External benefits may also flow from personal health-promoting activities, such as inoculation and vaccination, suggesting that people should perhaps be paid to undergo such prophylactic treatments so as to reflect the benefits they confer on others. It has, however, been more common to resort to compulsion, since it takes only a few 'non-co-operators' to undermine the benefits to the co-operators (such procedures seldom give 100% protection), and it must be acknowledged that the elimination of smallpox is a major triumph for that approach, making it now unnecessary for anyone to be vaccinated.

Distributional Judgements

Smallpox eradication was essentially a case where costs were imposed on some, in order that others might benefit, and in this respect it seems similar to the vexed case of fluoridation of water supplies, where the beneficial effects on young people seem beyond doubt, but the most efficient way of getting the fluoride into their teeth is by increasing the levels of fluoride in most public water supplies. But this means that the fluoride has to be consumed also by people who will get no benefit, and there may even be a few who may be harmed (as with vaccination). A distributional judgement has then to be made, and it cannot be escaped by *actual* compensation of objectors, because for some of them it seems to be an ethical issue of fanatical proportions, and they are never going to be 'bought off'.

Who Is the Best Judge of My Welfare?

For economists generally the toughest line of justification for public health measures of a coercive kind is the argument that consumers are not the best judges of their own welfare. Our strong upbringing in notions of consumer soverignty, and in the individualist calculus of the market, makes it very hard for us to accept that there are such things as ‘merit goods’ or ‘demerit goods’, i.e. goods where we think that someone other than the consumer is the best judge of whether that good is good or bad *for the consumer himself* (i.e. this has nothing to do with externalities, though they may also be present). We may be able to accept it for children, the mentally handicapped, or the mentally ill, but why should anyone (except by dependents) insist that I wear a seat belt when a passenger in a car, since only my own health is at stake? It may be argued that the costs of treating me (or burying me) will fall on others, in which case perhaps I should be forced to pay higher insurance premia if I wish to behave in that way. Is the argument then that there are some risks which we are not going to *permit* people to take, because we are confident that they will subsequently regret it. This line of reasoning seems strong in the case of the wearing of crash helmets by motorcyclists, who are mostly young people who seem to place ‘too low’ a value on their own lives and health.

Consumer Sovereignty?

Which brings us to the major contemporary public health issue, smoking. Assuming that smokers are compos mentis, well-informed and that the taxes they pay on tobacco fully cover any additional costs they impose on public services, and that they never smoke in any circumstances which place other people at risk (including the fire risks). There is presumably pleasure to be obtained from smoking, as there is from drinking, gambling, going to the opera, or playing bridge. Why then should we pressurize people to stop smoking, or never to start? The only reason would appear to be its addictive characteristics, i.e. there is a kind of ‘ratchet’ effect on the demand function for smoking, making it much more responsive upwards when prices fall or incomes rise, than it is downwards when prices rise of incomes fall. Moreover, smokers themselves seem to acknowledge the difficulty of ‘giving up’ on a scale that exceeds that of drinkers or gamblers or opera-goers or bridge players. Addictive behaviour sits awkwardly alongside consumer sovereignity, and especially when the producers of addictive substances are devoting large sums of money to persuading consumers to experiment with their products. It is an area of public policy where the ‘market failure’ and ‘public choice’ literatures fuse into a most excruciating scenario of conflicting ideologies and interest groups, with the economics of public health caught up in the difficulty of not knowing quite how much weight to give to the pleasure of smoking in such a tangled situation.

Public Health as an Investment

So let us come back to prevention being better than curve. We have rehearsed various phenomena (consumer ignorance, externalities, free-rider problems, economies of scale, and consumer incompetence) which have been adduced as reasons why there is a role for ‘public health’ measures of a preventative kind for people who are still well, alongside personal or state measures of a curative kind for people who are already ill. This leaves only one other consideration to be stressed, namely that many such preventative actions involve interventions at time t, which will not generate benefits until $t + n$, where n may be measured in decades rather than days or months. Thus a justification in *principle* for public health interventions may not translate into a justification *in practice*, because it needs to be subjected to empirical investigation to see whether the rate of return is worth it. Here the ‘rate of return’ will be measured as an improvement in the present value of people’s future health per unit of present value of the resources used, which has to be better than that obtainable from any alternative use of the resources. It is here that preventative programmes frequently show up badly. It seems persuasive to argue that a particular screening test costs only the equivalent of a packet of cigarettes, and it will save a hundred lives a year. But it frequently turns out that the costs of the tests are severely understated (e.g. ignoring the costs borne by patients in getting to and from where the test is given), that the costs of follow-up tests for ‘positives’ are excluded, as are treatment costs, both of which may be incurred (ineffectively) for many more people than the 100 people whose lives are ‘saved’ (i.e. whose death is ‘deferred’ . . . for how long? and with what quality of life?). And what if it takes a million tests to detect one successfully treatable case? And what discount rate do we use to reduce both costs and benefits to ‘present values’? Even a rate as low as 2% can have a devastating effect on benefits that are not going to show up for 40 years, as might well be the case with any public health intervention which would stop teenagers from smoking.

Public Health and the Economy

Thus it is that economists working in this field (as with many other areas of public policy) have not won enthusiastic acclaim from the professionals, or indeed from the public, who suspect that we are obsessed with income issues, such as maximizing conventionally measured GNP by making sick workers better as soon as possible, that we have lost sight of the broader issues, and especially of the fact much illness and injury is in fact *caused* by the unthinking pursuit of short term profit in activities that are good for GNP (as conventionally measured) but bad for human welfare on any commonsense interpretation. I have much sympathy with this broad viewpoint, and we need to do a lot more work on the extent to which economic activity produces ill health as well as goods and services for people to buy. But nothing in what I have written here is in any way affected by that important observation. Perhaps in the next issue of Palgrave there will have been enough work done on *this* issue for there to be a place for an entry on ‘The Economy as a Producer of Ill Health’, as well as on the Economics of Health, the Economics of Public Health, the Economics of State Provision of Medical Care, and so on.

See Also

- ▶ Health Economics
- ▶ State Provision of Medical Services

Bibliography

Atkinson, A.B.., and J.L. Townsend. 1977. Economic aspects of reduced smoking. *Lancet* 492–495.

CIBA Foundation. 1985. *The value of preventive medicine*. London: Pitman.

Jones-Lee, M.W. 1982. *The value of life and safety*. Amsterdam: North-Holland.

Leu, R., and T. Schaub. 1984. Economic aspects of smoking. *Effective Health Care* 2: 111–122.

Mooney, G. 1977. *The value of human life*. London: Macmillan.

Russell, L. 1986. *Is prevention better than cure?* Washington, DC: Brookings.

Wikler, D.I. 1978. Persuasion and coercion for health: Ethical issues of government efforts to change lifestyles. *Millbank Memorial Fund Quarterly* 56: 303–338.

Public Infrastructure

Teresa Garcia-Milà and Therese J. McGuire

Abstract

Numerous empirical studies have investigated the contribution of public infrastructure (the stock of publicly provided physical capital) to private economic productivity and growth. Using aggregate time-series data to estimate a production function with private capital, labour and public capital as inputs, the authors found substantial elasticities of private output with respect to public infrastructure. The result did not withstand scrutiny. Studies using disaggregated data (by region and industry), employing econometric diagnostics testing for nonstationarity, fixed effects and endogeneity, and using natural-experiment techniques found public infrastructure’s contribution to economic growth to be minor.

Keywords

Aggregate production functions; Cobb–Douglas functions; Cost functions; Output elasticity of capital; Productivity growth; Public infrastructure

JEL Classifications

H5

Public infrastructure (the stock of publicly provided physical capital comprising highways, sewage and sanitation systems, water systems, school buildings, hospitals and so forth) comprises an important component of the US economy. In 2004, there were just under seven trillion dollars of public capital in the United States, including 1.7 trillion dollars of highways and streets. By comparison, the stock of private capital stood at 27 trillion dollars.

Public infrastructure has figured in several related economic enquiries. What are the causes of private-sector productivity growth and decline? What causes lesser developed regions or countries to grow? Do country growth rates tend to converge over time? Research into each of these questions has examined the role of public infrastructure.

The contribution of public infrastructure to economic productivity and growth has been the focus of many empirical studies in recent years. The idea that public infrastructure should be considered an input in the aggregate production function, together with labour and private capital, was introduced in early theoretical models, but it was not taken into account in empirical work until the late 1980s. The increased attention to empirical analysis was linked to concerns about the decrease in productivity observed in the United States after 1970. In Europe much of the infrastructure literature has examined the role of public investment in boosting the growth of less developed regions.

The Theoretical Framework

The most widely used framework for studying the impact of public infrastructure on productivity and growth has been estimation of aggregate production functions where public capital is considered a production input along with the standard inputs, labour and private capital.

The general form of the aggregate production function can be written as follows:

$$Y_{rt} = A_{rt}F(L_{rt}, Kp_{rt}, Kg_{rt})$$

where Y is a measure of output, L represents labour, Kp is the stock of private capital, Kg is the stock of public capital, A is total factor productivity and the subscripts allow for regional and time variation.

Although the production function can take many functional forms, most empirical studies have estimated a Cobb–Douglas production function. Under that specification, and taking a logarithmic transformation, the estimating equation can be written as follows:

$$y_{rt} = a_{rt} + \alpha kp_{rt} + \beta kg_{rt} + \gamma l_{rt} + \varepsilon_{rt}$$

where the variables are measured in natural logarithms, and ε is an error term.

The aim of an important part of the public infrastructure literature is to estimate the output elasticity of public capital β, so as to measure its contribution to private productivity. An alternative approach to obtaining similar parameters, based on the duality of production and cost functions, is to estimate a cost function.

The Empirical Evidence

The first widely known results were those obtained by Aschauer (1989), who estimated a production function using aggregate post-war time series data for the United States. He estimated an output elasticity of public capital of 0.39, larger than the corresponding value for private capital (0.35). These estimates imply large returns for public investment (above 60 per cent, double of those for private capital), and were challenged by many authors, who considered them implausibly high.

Some authors attributed the high output elasticity of public capital found by Aschauer (and by Munnell 1990a) to a spurious correlation between output and public capital due to a common time trend. Aaron (1990) and Tatom (1991) corrected for the common time trend by first-differencing the data and obtained small and statistically insignificant coefficients.

The incorporation of state or metropolitan level data adding cross-section information to the time-series data opened up new possibilities for handling the spurious correlation and getting more accurate estimates of the contribution of public capital. Eberts (1986) focused on the manufacturing sector for 38 metropolitan areas and obtained a significant but much smaller estimate of the output elasticity of public capital, with a value of 0.03. Munnell (1990b) and Garcia-Milà and McGuire (1992) pooled state and time variation and obtained public capital elasticity estimates that range between 0.04 to 0.15. These estimates struck researchers as being more reasonable; however, they were subject to criticism because of endogeneity problems related mainly to omission of state-specific characteristics and to reverse causality.

Holtz-Eakin (1994), Evans and Karras (1994), and Garcia-Milà et al. (1996) used panel-data techniques not only to take into account state-specific productivity differences in the estimation, but also to explore non-stationarity of the data and possible endogeneity of the production factors. In all cases the estimates of the output elasticity of public infrastructure dropped dramatically compared with the time-series and pooled-data estimates, with values close to zero (and sometimes even negative) and not statistically significant.

By disaggregating by industry, Fernald (1999) avoided the endogeneity problem: if road infrastructure grew as a result of overall economic growth, and therefore the causality were reversed, one would not expect to find a relationship between increases in road infrastructure and the productivity of some industries but not others. He found that an increase in road infrastructure enhances productivity growth of vehicle-intensive industries much more than other industries. Fernald concluded that the US interstate building of the late 1950s and 1960s was one important factor in explaining the productivity increases up to the early 1970s, but the impact of road-building after the main network was built was small. A productivity burst because of road-building is a one-time effect and cannot be historically repeated. This is also the view of Hulten and Schwab (1993), who argue that, once the basic network is constructed, which has a major impact on the economy of the country, additional road construction has little, if any, effect on private productivity.

The results of Mas et al. (1996) support the idea that the impact of infrastructure investment is greater at earlier stages of development of the infrastructure network. They examine regions in Spain over the period 1964–91 and find that the output elasticity of productive infrastructure is 0.14 in the first ten years of the sample, but falls as more recent years are added to the sample, with a value of 0.08 when the whole period is considered. As the highway network in Spain was not yet completed in 1991, their results for Spain are compatible with those obtained for the United States, where the highway network was virtually complete, which showed that highway construction produced little or no effect.

Another possible response to the endogeneity problem is to estimate aggregate cost functions. The estimation of aggregate cost functions avoids the endogeneity bias if one can assume that prices of inputs are exogenous. Although it is reasonable to assume that individual firms are input price takers, it may not be plausible to assume that input prices are exogenous when considering aggregate (state or national level) cost functions. In spite of this shortcoming, there are several interesting papers that estimate aggregate cost functions and obtain, through the duality between cost and production functions, estimates of the output elasticity of private and public inputs. Berndt and Hansson (1992), Lynde and Richmond (1992), Nadiri and Mamuneas (1994), and Morrison and Schwartz (1996) are good examples of cost function estimations. They differ in the functional specification, the geographical and industrial scope, and the scope of public infrastructure, but in all cases they find that public capital reduces costs and therefore improves productivity. The size of the effect tends to be quite small, along the lines of the production-function estimates obtained by Eberts (1986), Munnell (1990b) and Garcia-Milà and McGuire (1992).

A very different way to avoid endogeneity bias is to look for a natural experiment related to infrastructure investment. The idea is to compare the economic performance of two otherwise similar areas except that in one area, the control group, there has not been any highway construction whereas the other area, the treated group, has experienced highway construction. Rephann and Isserman (1994) apply matching techniques to analyse the effectiveness of US interstate highways as an economic development tool. Results vary significantly depending on the characteristics of the counties considered. Counties close to a large city or containing small cities (more than 25,000 residents) benefit from new investments in interstate highways, while rural counties without these characteristics do not experience economic growth when interstate highways are built within them. Chandra and Thompson (2000) exploit the fact that in the United States much interstate highway construction was designed to link major metropolitan areas. Thus, the rural counties in between the metropolitan areas through which the interstates run can be considered the treatment group, while the counties adjacent to the treatment group, which in essence just missed having an interstate highway run through them, can be considered as a suitable control group. The authors find that counties that have a new interstate highway running through them experience an increase in overall earnings, while earnings fall in the adjacent counties. The authors conclude that interstate highway construction affects the spatial allocation of economic activity, but has no net effect on the economic development of non-metropolitan areas as a whole.

The question of whether investment in public capital yields a net positive social return has also been addressed. Morrison and Schwartz (1996) calculate a measure of the net social return to public infrastructure investment as the difference between the cost savings to manufacturing firms minus the cost to society of the public capital investment. Their results range from small positive values to negative estimates of the net social return depending on how the price of public capital is adjusted for taxation and for the marginal cost of public funds. Haughwout (2002) examines the impact of public infrastructure on both productivity and consumer utility in a sample of large US cities. He finds that the local benefits of public capital are largely realized by households rather than firms and that the aggregate benefits of large investments in public infrastructure are not likely to be sufficient to offset the costs.

Concluding Remarks

Based on the aggregate analysis, we can conclude very little. The most credible aggregate production–function estimates of the impact of public infrastructure on private output hover around zero, as do estimates of the net social benefit of public infrastructure investment. However, when the focus is on sub-aggregates such as particular industries or certain areas or incomplete networks, researchers tend to find that public capital investment boosts private output and productivity for some industries, some areas, and some networks. What is clear from the accumulated evidence is that public infrastructure is not a panacea for all that ails economies, but rather a tool that when properly targeted can be effective at enhancing growth.

Bibliography

Aschauer, D.A. 1989. Is public infrastructure important? *Journal of Monetary Economics* 23: 177–200.

Aaron, H.J. 1990. Discussion of ‘why is infrastructure important?’. In *Is There a Shortfall in Public Capital Investment?* ed. A.H. Munnell. Boston: Federal Reserve Bank of Boston.

Berndt, E.R., and B. Hansson. 1992. Measuring the contribution of public infrastructure capital in Sweden. *Scandinavian Journal of Economics* 94(Supplement): 151–168.

Chandra, A., and E. Thompson. 2000. Does public infrastructure affect economic activity? Evidence from the rural interstate highway system. *Regional Science and Urban Economics* 30: 457–490.

Eberts, R.W. 1986. Estimating the contribution of urban public infrastructure to regional growth. Working Paper No. 8610, Federal Reserve Bank of Cleveland.

Evans, P., and G. Karras. 1994. Are government activities productive? Evidence from a panel of U.S. states. *Review of Economics and Statistics* 76: 1–11.

Fernald, J.G. 1999. Road to prosperity? Assessing the link between public capital and productivity. *American Economic Review* 89: 619–638.

Garcia-Milà, T., and T.J. McGuire. 1992. The contribution of publicly provided inputs to states’ economies. *Regional Science and Urban Economics* 22: 229–241.

Garcia-Milà, T., T.J. McGuire, and R.H. Porter. 1996. The effect of public capital in state-level production functions reconsidered. *Review of Economics and Statistics* 78: 177–180.

Haughwout, A.F. 2002. Public infrastructure investments, productivity, and welfare in fixed geographic areas. *Journal of Public Economics* 83: 405–425.

Holtz-Eakin, D. 1994. Public-sector capital and the productivity puzzle. *Review of Economics and Statistics* 76: 12–21.

Hulten, C.R., and R.M. Schwab. 1993. Infrastructure spending: Where do we go from here? *National Tax Journal* 46: 261–273.

Lynde, C., and J. Richmond. 1992. The role of public capital in production. *Review of Economics and Statistics* 74: 37–44.

Mas, M., J. Maudos, F. Perex, and E. Uriel. 1996. Infrastructures and productivity in the Spanish regions. *Regional Studies* 30: 641–649.

Morrison, C.J., and A.E. Schwartz. 1996. State infrastructure and productive performance. *American Economic Review* 86: 1095–1111.

Munnell, A.H. 1990a. Why has productivity growth declined? Productivity and public investment. New England Economic Review, 3–22, January/February.

Munnell, A.H. 1990b. How does public infrastructure affect regional economic performance? In *Is there a shortfall in public capital investment?* ed. A.H. Munnell. Boston: Federal Reserve Bank of Boston.

Nadiri, M.I., and T.P. Mamuneas. 1994. The effects of public infrastructure and R&D capital on the cost structure and performance of U.S. manufacturing industries. *Review of Economics and Statistics* 76: 22–37.

Rephann, T., and A. Isserman. 1994. New highways as economic development tools: An evaluation using quasi-experimental matching methods. *Regional Science and Urban Economics* 24: 723–751.

Tatom, J.A. 1991. Public capital and private sector performance. *St. Louis Federal Reserve Bank Review* 73: 3–15.

Public Sector Borrowing

Terry Ward

Public sector borrowing is the difference between the total receipts of central government, local authorities and public enterprises, considered in aggregate, and their total expenditure. As such it is somewhat broader in coverage than the public sector financial balance or the Federal deficit in the US which largely, though not entirely, exclude financing items. It thus encompasses the issue of short and long-dated securities not only to finance government expenditure of the conventional kind on goods and services and transfers, but also to fund investment projects carried out by public enterprises, which might yield attractive commercial rates of return. It also includes purely financial transactions such as lending to the private sector or purchases of assets, typically treated as 'below-the-line' items and as part of the Credit Budget in the US.

Each of these types of activity has very different implications for aggregate demand, real output and monetary conditions in the economy. Purely financial transactions in particular may have very little macroeconomic relevance at all. This together with the fact that borrowing relates to the net cash flow position of the public sector rather than to the gap between income and outlays measured on an accruals basis, means that the magnitude is liable to give a misleading impression of its financing implications as well as of its wider influence on the economy. Moreover it does not even give a reliable indication of the total value of securities which the government needs to sell over any particular period to fund its activities, since this will be affected by the amount of existing debt which has to be rolled over as well as by the monetary policy being followed (sales of government debt may be used as a means of controlling credit or the level of interest rates). Accordingly it ought to be magnitude of relatively little interest to economists and much less relevant than the public sector financial deficit or budget deficit which is largely confined to 'above-the-line' items and excludes financing transactions.

These shortcomings have, however, not prevented public sector borrowing becoming a focal point of government policy in number of countries, including the UK. A major reason for this is the rise of monetarism and the central role assumed by monetary targets in the conduct of macroeconomic policy since the mid-1970s throughout the industrialized world. Since public sector borrowing is an important component of money supply growth and the part ostensibly under the most direct control of government, it has come to be regarded as the key to keeping monetary growth within the limits thought to be necessary for containing inflation, without putting undue upward pressure on interest rates.

In practice, things are not quite so simple. Attempts to control public sector borrowing by raising tax rates or cutting public expenditure programmes are liable to have depressing effects on economic activity and real incomes. This in turn can accordingly give rise to a greater need for private sector borrowing to maintain expenditure and, in the case of companies, to avoid liquidation. A reduction in one major component of monetary growth can therefore induce a compensating rise in the other major component. This helps to explain why, empirically, there appears to be no close correlation between public borrowing and money supply growth in many countries.

Nevertheless, imposing limits on public sector borrowing or on the budget deficit represents an ultimately effective and readily verifiable, if crude means of reducing financial instability and containing national debts, even though it may well be at the cost of real output growth and employment. On the other hand, the fact that public borrowing, and to a less extent the budget deficit, include some transactions which have relatively little effect on economic activity gives some scope for governments to manipulate its scale and thereby appease financial markets without markedly depressing output or real income. Such so-called ‘window-dressing’ was a feature of the interwar years and has re-emerged in the UK, for example, in the 1980s. Ironically, however, the scope for such manipulation tends to be greatest for right-wing market-oriented Governments which are relatively well disposed towards selling of State assets but which are usually least favourably inclined towards fiscal expansion.

See Also

- ► Burden of the Debt
- ► Deficit Financing
- ► Functional Finance
- ► Public Debt
- ► Ricardian Equivalence Theorem

Public Sector Investment Efficiency in Developing Countries

Alvar Kangur and Chris Papageorgiou
International Monetary Fund, Washington,
DC, USA

Abstract

There are numerous examples where public investment has been grossly mismanaged and where corruption has overwhelmed the entire process (unfinished roads, highways leading to nowhere, incomplete or unusable bridges and power generation projects). This entry aims at reviewing the existing literature on the potential impact of such public investment inefficiencies on productivity and output, in theoretical models and empirical exercises. We conclude that despite recent progress in assessing and incorporating such inefficiencies in economic analysis, the composition of public capital and its interlinkages with other factors of production and with structural economic conditions should remain a key area of future research.

Keywords

Public Investment; Marginal Product of Capital; Capital Stock; Government Inefficiencies

JEL Classification

O19; O23; O47

Unlike with private investors, there is no plausible behavioral model in which every dollar that the public sector spends as “investment” creates economically valuable “capital”. While this simple analytic point is obvious, it has so far been uniformly ignored in the empirical literature on economic growth. Lant Pritchett (2000)

The views expressed in this project are the sole responsibility of the authors and should not be attributed to the International Monetary Fund, its Executive Board, or its management.

Introduction

There is a broad consensus that a scaling-up of investment in developing countries, particularly in infrastructure, is critical to achieve sustained growth. Particularly in many low-income countries, deficiencies in infrastructure may sometimes reduce productivity by at least as much as structural factors, such as bureaucracy, corruption and lack of financing. Recent studies by the World Bank and the IMF suggest that the growth impact of higher infrastructure spending in low-income countries is potentially substantial – if low-income countries halved their infrastructure gaps, reaching the level of middle-income countries, annual growth rates would increase by about 2% points.

In many developing and low-income countries, however, the link between public capital spending and capital stock accumulation, and hence growth, is weakened by evidence of low efficiency of public investment. The notion that public investment spending leads to equivalent capital accumulation rests on the assumption that public investment is inherently productive. This assumption is particularly problematic in many developing countries, as a high degree of inefficiency, waste, or corruption often distorts the impact of public spending on capital accumulation, leaving a trail of poorly executed and ineffective projects.

While the literature suggests that a scaling-up of investment in developing countries is vital, the link with outcomes depends critically on the quality and efficiency of public investment. This highlights the importance of going beyond discussions of spending levels and addressing issues of the broad institutional framework underpinning the provision of investment. As accurately described by Caselli (2005) "less-accountable poor-country governments are likely to be disproportionately less efficient (relative to the private sector) than rich country ones. Hence, there are good reasons to expect the government to play an especially detrimental role in the productivity of investment in poor countries". This translates into variability in the market value of the capital stock.

In terms of theory, this would imply adjusting our standard models of economic growth. A clear way to achieve this would require incorporating in the capital accumulation process, a parameter to capture public investment inefficiency, and consequently modifying the stock of public capital in the aggregate production function. Recognizing that both the investment and the stock of public capital are compromised by government inefficiencies is likely to alter predictions of standard models. In terms of empirics, being able to estimate the difference between investment cost and capital value is of first-order empirical importance especially for developing countries for whom public investment is the primary source of investment. In practical terms, assessing the quality of project selection, appraisal, implementation and evaluation in a country can help identify the specific weaknesses that contribute to poor outcomes and guide appropriate institutional and technical remedies that could correct such failures.

This entry is structured as follows: In section "Two Generations of Research on Public Capital" we provide a selected literature review focusing on two distinct generations of analytical work on the productivity of public capital. In section "Recent Work Sheds More Light on the Way that Public Investment", we turn our attention to estimation issues, particularly related to efficiency of public capital, by challenging some of the main assumptions made in the basic growth models. Subsequently, we consider and discuss in some detail different measures of inefficiencies found in the literature, and conclude the section with some examples of empirical applications using such measures. In section "Towards the Third Generation of Research" we present recent work that focuses on general equilibrium models that incorporate public investment inefficiencies and analyse in detail their effects not only on aggregate output but also several sectors of the economy. Finally, section "Conclusions" concludes with discussion on future research.

Two Generations of Research on Public Capital

Substantial research has been devoted to measuring the productivity of public capital. In response to the massive public sector investment booms in many developing countries in the 1970s, Little and Mirrlees (1974) provided a systematic and practical cost-benefit methodology to assess public investment decisions. In a follow-up article, Little and Mirrlees (1990) expanded on their methodology and demonstrated some of the benefits of their approach which inspired an extensive literature that lasted for decades and is the starting point of most project evaluation work today. Partly using some of the work by Little and Mirrlees, the first generation of research that started in the late 1990s typically found that public capital can offer very large productivity gains, notwithstanding the wide range of theoretical and empirical frameworks employed. Aschauer (1989, 1998) in a series of papers estimated the output elasticity of public capital in the range of 0.3–0.4 and was the first to assign public capital an important role in explaining the fall in US productivity growth observed in the 1970s and 1980s. The literature that followed largely confirmed Aschauer's findings. Munnell (1990a, 1992) estimated the impact of public capital on growth at 0.31–0.39 at the national level though in Munnell (1990b) found a lower impact of 0.15 at the state level. In a similar setting, Lynde and Richmond (1993) found that the services of public capital are an important part of the production process and that about 40% of the slowdown in the growth rate of labour productivity is explained by a fall in the public capital-labour ratio. Several other papers reached similar conclusions; see Sturm et al. (1998), for a comprehensive review of this generation of studies.

Over time these first-generation estimates were questioned on the grounds of numerous methodological and econometric limitations (Gramlich 1994). Issues ranking high on the list of potential problems included reverse causation from productivity to public capital (public capital affects productivity, and in turn is affected by productivity) and spurious correlation due to non-stationarity (time-varying properties of the public capital series). This controversy sparked a new generation of research, which compared to the results surveyed by Sturm et al. (1998) estimated substantially lower effects of public capital on growth; see Romp and de Haan (2007) for an extensive review. Moreover, while attempting to address the aforementioned estimation problems, the research unveiled substantial heterogeneity among countries, regions and sectors. This is not surprising, as the effects of new investment spending depend on the quantity and quality of the capital stock in place. In general, the larger the stock and the better its quality, the lower will be the impact of every additional unit of capital added to this stock (the marginal productivity of capital). The network character of public capital, notably of infrastructure, also results in non-linearities in the impact of public capital on growth. It is these non-linearities which explain some of the above heterogeneity. Thus, the effect of new capital will crucially depend on the extent to which investment spending is targeted to alleviate bottlenecks in the existing network. Further studies suggest that the effect of public investment spending on growth may also depend on institutional and policy factors (Tanzi and Davoodi 2000; Sawyer 2010).

Bom and Ligthart (2010) summarized the estimates of the output elasticity of public capital available from the literature by means of a meta-regression analysis. They find that the unconditional average output elasticity of public capital centres around 0.15 but suggest substantial heterogeneity across countries. They also show that studies that impose constant returns to scale restrictions across private labour and capital (Mas et al. 1993; Otto and Voss 1994; Kavanagh 1997), control for the business cycle (Aschauer 1989; Hulten and Schwab 1991; Sturm and De Haan 1995), and incorporate some measure of education (Garcia-Milà and McGuire 1992) find larger output elasticities of public capital, whereas studies that include energy prices (Tatom 1991) tend to find lower estimates.[1] The conditional output elasticity of public capital in

[1] Imposing constant returns to scale across private inputs implies increasing returns to scale across all inputs if the factor share of public capital is positive. This could produce upward bias in the estimates if the true model is characterized by decreasing returns to scale across private inputs.

the regression equation which captures typical study characteristics is estimated at 0.17, which is not that far from its unconditional (without capturing the study characteristics) value of 0.15. These values imply a marginal productivity of public capital for the United States in the range of 29–33% in 2001.

Given data limitations and the difficulty in constructing public capital stock series for developing countries, the early empirical literature on these countries often looked directly at the impact of public investment on economic growth (Devarajan et al. 1996). Arslanalp et al. (2010) was among the first to estimate a production function using the public capital stock as an explanatory variable, for a sample of 48 developed and developing countries. The effect of public capital on growth is estimated to be stronger for developed countries in the short term (0.13), while it is stronger for developing countries in the long term (0.26). In some countries, they find that the positive impact of public capital on output is partially or wholly offset if the initial ratio of the capital stock to GDP is high. Their results also show that in developing countries certain types of constraints (financing or the ability to absorb) can limit the growth benefits of higher capital stock and, unlike in advanced countries, the benefits of new investment tend to accrue more gradually.

Recent Work Sheds More Light on the Way that Public Investment

Efficiency might decline during investment booms. Warner (2014) looks at big long-lasting drives in public capital spending in developing economies, and concludes that only a weak positive association exists between investment spending and growth, and that too only in the same year. According to the author public investment drives have tended to be financed by borrowing and have been plagued by incentive problems and interest-group-infested investment choices at the time investment projects were chosen. In addition to the inefficiency issues public investment booms are also faced by severe absorptive capacity issues. Presbitero (2016) reports evidence in a panel of a large number of developing economies that investment and infrastructure projects are less likely to be successful when they are undertaken during periods of higher than average public investment. This evidence is consistent with the presence of supply bottlenecks and poor project selection and with the importance of sound policies and institutions for the selection and management of public investment projects.

One of the most basic dynamic equations in macroeconomics is that of the accumulation of capital which (under some conditions) is given by:

$$G'_{it} = G'_{it-1} - \delta_{it}{}^{*}G'_{it-1} + q_i{}^{*}I_{it-1} \tag{1}$$

where for each country i, G'_t is the stock of public capital at time t, and I_{t-1} is public investment spending at time $t-1$. δ_{it} is country i's time-varying rate of depreciation of the capital stock. Equation 1 indicates that the stock of physical capital in any period is equal to the fraction of total investment converted into capital in addition to the existing undepreciated capital stock.

As most prominently noted by Pritchett (2000) – a criticism that goes to the heart of this topic – the behavioural model embedded in this universally used equation and hence in all of the existing empirical literature as summarized previously assumes full public-sector efficiency (i.e. $q_i = 1$) even when there are no empirical or theoretical grounds for making such an extreme assumption. To the contrary, it is widely believed that in many countries only a fraction of the actual accounting cost of investment passes into the value of capital. Yet this obvious point is routinely ignored and cross-national estimates of physical capital still continue to be based on the assumption of full efficiency of public investment. As such, the assumption of full public capital efficiency cannot be the last act in drawing meaningful conclusions on the impact of public capital or investment on growth.

One does not have to look far to see the difference the public investment efficiency can make. The quality of infrastructure component of *The Global Competitiveness Index* taken from *World Economic Forum* (2016) shows that for

Public Sector Investment Efficiency in Developing Countries, Table 1 Public Investment Management Index (PIMI) by income group

	PIMI	Appraisal	Selection implementation		Evaluation
Low income (40)	0.47	0.21	0.28	0.30	0.20
	(0.26)	*(0.13)*	*(0.11)*	*(0.10)*	*(0.10)*
Middle income (31)	0.57	0.21	0.30	0.28	0.22
	(0.25)	*(0.09)*	*(0.11)*	*(0.07)*	*(0.07)*
All countries (71)	0.51	0.21	0.29	0.29	0.21
	(0.26)	*(0.11)*	*(0.11)*	*(0.09)*	*(0.09)*

Sources: Dabla-Norris et al. (2012) and authors' calculations. Standard deviations are in parentheses.

advanced countries the quality of infrastructure measured on the 1–7 scale is clustered at the high end with very low variation, suggesting that efficiency differences might not play a pivotal role among these countries. However, for low- and middle-income countries the median scores are almost two points lower than advanced countries with much larger variation in the scores. Thus, to truly understand the impact of investment or capital efficiency on economic outcomes, it is most useful to explore this issue for the developing world.

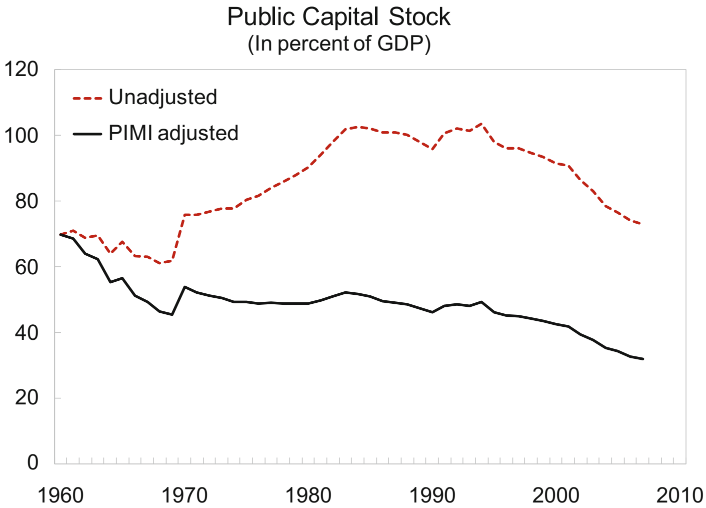
While the WEF index captures *the quality of capital*, for a researcher it is more informative to understand *the quality of the process* that turns investment into public capital. To our knowledge the first such more comprehensive Public Investment Management Index or PIMI is provided by Dabla-Norris et al. (2012) for the four stages of public investment management – appraisal, selection, implementation, and evaluation – covering 71 developing countries (40 low-income and 31 middle-income countries).[2] Table 1 suggests that, on average in the PIMI sample, only about half of public investment efforts translate into actual productive public capital. Even when accounting for possible biases that may exaggerate this finding, inefficiencies in public investment remain massive and are well recognized by both academics and policymakers alike.

This masks significant heterogeneity between countries and, even more notably, between each of the sub-indices. Further, the pair-wise rank correlations between PIMI and other similar indices such as the Budget Institution index constructed by Dabla-Norris et al. (2010), Kaufmann and Kraay (2008) governance indicators (including Government Effectiveness, the average of the Governance Indicators, and the Control of Corruption index) and the World Bank's (2009) Country Policy and Institutional Assessment (CPIA) index are positive though not overly high, ranging from 0.3 to 0.6. This indicates that the PIMI can carry information on the quality of public investment not fully captured by other more general institutional and governance indices and can thus be considered as a complement to, and not a substitute for, these more general indices.

Gupta et al. (2014) are the first to directly adjust public investment for efficiency. Their methodology to accumulate the capital stock series by Eq. 1 is similar to that used by Collier et al. (2001), Kamps (2006) and Arslanalp et al. (2010). For the crucial efficiency parameter q_i they use normalized PIMI as well as its four subcomponents. The Fig. 1 showing the general results of capital accumulation exercise indicates a significant gap between the traditional and efficiency-adjusted public capital stock in the order of 40% of GDP in the recent years available. It is also remarkable that throughout the sample period, and contrary to the unadjusted stock, efficiency-adjusted capital has substantially declined – a trend led mostly by low-income

[2] To construct PIMI data were compiled from a large number of sources including from World Bank Public Investment Management case studies, Public Expenditure and Financial Accountability assessment reports, the Budget Institutions database, World Bank Public Expenditure Reviews, World Bank Country Procurement Assessment Reviews, World Bank Country Financial Accountability Assessments and country websites.

Public Sector Investment Efficiency in Developing Countries, Fig. 1 Public Capital Stock (In percent of GDP)



countries – indicating that high-quality public investment would be associated with a high marginal product.

Armed with the new efficiency-adjusted capital stocks Gupta et al. (2014) proceed to estimate an otherwise standard unconstrained Cobb-Douglas production function:[3]

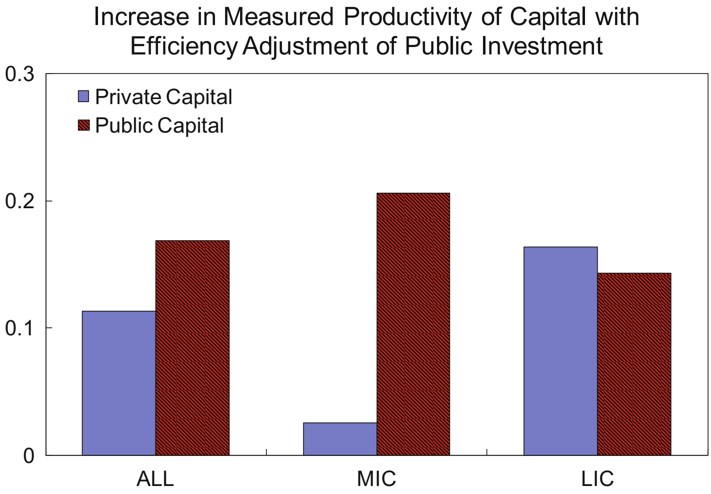
$$Y_{it} = A_0 S_{it}^{\alpha} K_{it}^{\beta} G_{it}^{\gamma} e^{\lambda_t + \varepsilon_{it}} \qquad (2)$$

where skill-adjusted labour S_t is computed according to $S_{it} = L_{it}{}^{*} e^{\varphi(h)}$, where L_{it} is raw labour and h is the average years of schooling in the population aged 15 years and older. $\varphi(h)$ is a stepwise linear function adjusting the average years of schooling by estimates for returns on education. The econometric results indicate that adjusting public capital for public investment efficiency better explains the evolution of relationship between public capital and growth. The efficiency adjustment reduces the estimated share of public capital in the production function to around 0.15 that is statistically significant for both low- and middle-income countries. More importantly, it leads to a corresponding increase in the shares of private capital, especially for LICs. As a result of efficiency-adjustment and corresponding changes in estimated production function coefficients, the marginal productivity of both private and public capital increases (see Fig. 2). The increase in private capital productivity is higher in low-income countries (LIC), whereas the increase in public capital productivity is higher in middle-income countries (MIC).

PIMI components can be omitted one-by-one, from the accumulated stock of public capital. This exercise indicates that the importance of investment stages for productivity of public capital varies with income levels. Project implementation (which comprises competitive and open bidding and internal audit) is the most critical component of the investment process. This result, which holds on aggregate, is driven mostly by low-income countries in the sample, for whom project selection (that is related to medium-term framework) assumes secondary importance. For the middle-income countries, project appraisal (which comprises transparency of appraisal standards) and projection evaluation (which comprises external audits) are relatively more important. While for all countries for which PIMI is available, implementation stands out as the stage with higher relative productivity, the results for sub-samples are mixed. This indicates that new public investment must be accompanied

[3] Aggregate data on output, investment and raw labour are taken from PWT version 6.2, public investment shares to accumulate capital stocks are taken from WEO databases, and the average years of schooling come from the Barro and Lee (2013) database on educational attainment.

Public Sector Investment Efficiency in Developing Countries, Fig. 2 Increase in Measured Productivity of Capital with Efficiency Adjustment of Public Investment



by strengthening of specific bottlenecks in investment processes to enhance the productivity of public capital.

Towards the Third Generation of Research

A number of theoretical applications have directly modelled public sector investment efficiency as in Eq. 1 in the DSGE framework to allow for a richer set of interactions, including the presence of a zero lower bound on nominal interest rates. Berg et al. (2013) are among the first to model declining investment efficiency in the stock of public capital capturing capacity constraints. Especially for resource rich countries, this can support a more gradual public investment strategy for windfall savings that could initially be saved in an external fund.

More importantly, Berg et al. (2015) identify and clearly explain a key steady-state result characterized by the invariance of growth with respect to investment efficiency: the impact of additional investment on the growth rate of output does not depend on the level of the time-invariant efficiency parameter q. This is easy to see in a two-equation system consisting of capital accumulation Eq. 1 and a Cobb-Douglas production function such as Eq. 2 with only public capital as an input:

$$Y_t = A_t G_t^{\gamma}. \tag{3}$$

Equation 1 implies that, in a steady state:

$$K = \frac{qI}{\delta}. \tag{4}$$

The rate of return of a marginal unit of investment can then be expressed as follows:

$$\frac{dY}{dI} = \frac{dY}{dK} * \frac{dK}{dI} = \gamma \frac{Y}{I} \tag{5}$$

Equation 5 implies that the impact of a marginal change in investment on the growth rate of output (which is a product of the marginal productivity of capital and the capital stock per unit of investment) is simply equal to the public capital share γ in the production function and does not depend on the level of investment efficiency q. The general intuition to this invariance lies in the law of diminishing marginal productivity: since the time-invariant efficiency q permanently scales down the capital stock, diminishing returns imply higher marginal productivity. These two effects work in the opposite direction and with Cobb-Douglas exactly offset each other. This implies that inefficiency, per se, would not lead to lower growth and should not be considered as a reason to not invest.

However, it must be noted that the impact of public investment and its efficiency on output and growth can be more complex than implied by the aforementioned two-equation system. At least four key reasons can be brought out that can support growth effects of investment efficiency. First, the Berg et al. (2015) is a long-run *steady-state result* that does not preclude that changes in investment efficiency affect *transitional growth* towards the steady state, in a manner similar to the saving rate in exogenous growth models. Second, as is also acknowledged by the authors, patterns of complementarities between the production factors matter: if public and private capital are complements, public investment in low-efficiency countries could still have an impact on the growth rate of output through higher marginal productivity of private capital. Third, intuitively investment inefficiency can have an impact on growth (even at steady state) through scale effects, if the level of effective capital determines the level of technology at the aggregate level, creating positive externalities similar to the pioneer endogenous growth models of Romer (1986) and Lucas (1988). Fourth, the impact of investment and efficiency on growth can depend on the *structural conditions* of the economy, including cost of financing, the fiscal space, and the level of debt.

For example, Buffie et al. (2012) show that *productive* public and private capital are complements. However, if structural conditions are weak, including low public investment efficiency and collection rates, instead of crowding-in a surge in public investment can crowd-out private investment and could lead to unsustainable public debt. Model simulations by IMF (2014) show similar results for developing economies characterized by *structural conditions* that usually exhibit less slack, less accommodative monetary policies, and importantly lower public investment efficiency. In these economies, a public investment shock leads to substantially lower long-term output effects compared to advanced countries, and a higher public debt to GDP ratio that in turn can impinge on growth.

Finally, interactions between public and private capital pose an important policy question on the use of public-private partnerships (PPPs): if accounting for high public sector investment inefficiencies also leads to higher marginal productivity of private capital, can more widespread use of PPPs be associated with higher growth, especially in LICs? While research here is still scarce, it can prove to be a promising field. Buffie et al. (2016) document that, even if costlier, PPPs produce higher quality capital at shorter times compared to public sector own investment. Their general equilibrium simulations suggest that PPPs can have a social return 5–8 points higher than own public investment.

The work on the efficiency of public investment has highlighted several limitations and would benefit from research across the following dimensions. First, the PIMI is available only for one period 2007–2010 and is thus time-invariant. While it encompasses cross-sectional variation, it is not able to capture changes that have incurred to investment processes over time. Second, the determination of depreciation rates of public capital stock that vary across time and countries as well as the level of the initial public capital stock deserves further investigation to reduce measurement errors. Third, the empirical literature is yet rather silent on whether public capital is complementary to, or a substitute for, other production factors, including wealth of natural resources that is highlighted by Caselli (2005) as one of the factors that could bring the estimated marginal productivities of capital across countries closer together. Fourth, in an open-economy growth model with perfect capital mobility convergence would happen instantly as a fully integrated global economy ensures that differences in rates of return on capital are eliminated across countries. To explain why we do not observe this requires consideration of the possible frictions in international capital markets that slow down or eliminate convergence altogether. Obstfeld and Rogoff (1996) present an open-economy growth model that demonstrates in a very tenable and intuitive way the ability of market imperfections to yield convergence dynamics in an integrated global economy. Whether lack of efficient public capital can provide an explanation of why capital does not flow to less developed countries is a promising research avenue (see, e.g. Lowe, Papageorgiou and Perez Sebastian). Finally, investigation into the productivity of public capital would benefit from a wider exposure to different methodologies.

Direct estimation of the investment efficiency parameters in a DSGE model as done by Berg et al. (2013, 2015) is a useful alternative if reliable information is otherwise lacking.

Conclusions

Public investment in bridges, roads and ports is truly essential in low-income countries which suffer from massive infrastructure gaps. It is considered one of the most important drivers of growth and a primary component of the development strategies of governments in the developing world. Significantly boosting investment in physical infrastructure to achieve sustained growth rests on the high returns to investment in capital-scarce environments, and the pressing deficiencies in these areas. However, inefficiencies in project appraisal, selection, implementation and evaluation lead to devastating losses in public capital accumulation and output. The history of public investment booms is filled with disheartening stories about "roads to nowhere" and "white elephants" especially in poor countries where public goods are in dire need.

This entry suggests that considerable progress has been made by economists in better measuring, and more appropriately, incorporating public investment inefficiencies in economic models. In addition, our understanding of the drivers of these inefficiencies, such as poor incentive systems, inadequate capacity to appraise and implement projects, and absorptive capacity, has improved. Nonetheless, more needs to be done in this important area of economics. As better data become available, including at the firm and sectoral level, economists should improve existing indices of public investment inefficiencies. This is resource-intensive work that requires careful collaboration with governments and researchers. But without the necessary and high-quality data, assessment of inefficiency will not be possible. At the same time, country authorities must pay particular attention in improving their processes of project selection, appraisal, implementation and evaluation. This is a very attainable goal that encouragingly has started to become a priority in most economies of the developing world.

See Also

▶ Infrastructure and Growth
▶ Infrastructure and Inequality

Acknowledgments Papageorgiou acknowledges financial support from the U.K.'s Department for International Development (DFID).

Bibliography

Arslanalp S., F. Bonhorst, S. Gupta, and E. Sze. 2010. Public capital and growth. IMF Working Paper 10/175.

Aschauer, D.A. 1989. Is public expenditure productive? *Journal of Monetary Economics* 23: 177–200.

Aschauer, D.A. 1998. How big should the public capital stock be? The Jerome Levy Economics Institute of Bard College Public Policy, No: 43.

Barro, Robert, and Jong-Wha Lee. 2013. A new data set of educational attainment in the world, 1950–2010. *Journal of Development Economics* 104: 184–198.

Berg, A., R. Portillo, S.S. Yang, and L. Zanna. 2013. Public investment in research-abundant developing countries. *IMF Economic Review* 61 (1): 92–129.

Berg, A., E.F. Buffie, C. Pattillo, R. Portillo, A. Presbitero, and L. Zanna. 2015. Some misconceptions about public investment efficiency and growth. IMF Working Paper 15/272.

Bom, P, and J.E. Ligthart. 2010. What have we learned from three decades of research on the productivity of public capital? CESifo Working Paper Series No. 2206, Center Discussion Paper No. 2008–10.

Buffie E.F., A. Berg, C. Pattillo, R. Portillo, and L.F. Zanna. 2012. Public investment, growth, and debt sustainability: Putting together the pieces. IMF Working Paper 12/144.

Buffie, E.F., M. Andreolli, B.G. Li, and L. Zanna. 2016. Macroeconomic dimensions of public-private partnerships. IMF Working Paper 16/78.

Caselli, F. 2005. Accounting for cross-country income differences. In *Handbook of economic growth*, vol. 1, Part A, ed. P. Aghion and S.N. Durlauf, Elsevier B.V. 679–741.

Collier, P., A. Hoeffler, and C. Pattillo. 2001. Flight capital as a portfolio choice. *The World Bank Economic Review* 15: 55–80.

Dabla-Norris, E., R. Allen, L.F. Zanna, T. Prakash, E. Kvintradze, V. Lledo, I. Yackovlev, and S. Gollwitzer. 2010. Budget institutions and fiscal performance in low-income countries. IMF Working Paper WP/10/80, Washington, DC.

Dabla-Norris, E., J. Brumby, A. Kyobe, Z. Mills, and C. Papageorgiou. 2012. Investing in public investment: An index of public investment efficiency. *Journal of Economic Growth* 17 (3): 235–266.

Devarajan, S., V. Swaroop, and H.F. Zou. 1996. The composition of public expenditure and economic growth. *Journal of Monetary Economics* 37: 313–144.

Garcia-Mila, T., and T.J. McGuire. 1992. The contribution of publicly provided inputs to states' economies. *Regional Science and Urban Economics* 22: 229–241.

Gramlich, E.M. 1994. Infrastructure investment: A review essay. *Journal of Economic Literature* 32 (3): 1176–1196.

Gupta, S., A. Kangur, C. Papageorgiou, and A. Wane. 2014. Efficiency-adjusted public capital and growth. *World Development* 57: 164–178.

Hulten, C.R., and R.M. Schwab. 1991. Is there too little public capital: Infrastructure and economic growth. Conference paper, American Enterprise Institute, Washington, DC.

IMF. 2011. Fiscal monitor. Addressing fiscal challenges to reduce economic risks, Sept 2011.

IMF. 2014. Is it time for an infrastructure push? The macroeconomic effects of public investment. In *World economic outlook. Legacies, clouds, uncertainties*, Oct 2014.

Kamps, C. 2006. New estimates of government net capital stocks for 22 OECD countries, 1960–2001. *IMF Staff Papers* 53: 120–150.

Kaufmann, D., and A. Kraay. 2008. Governance indicators: Where are we, where should we be going? *World Bank Research Observer* 23: 1–30.

Kavanagh, C. 1997. Public capital and private sector productivity in Ireland. *Journal of Economic Studies* 24: 72–94.

Little, I.M.D., and J.A. Mirrlees. 1974. *Project appraisal and planning for developing countries*. New York: Heinemann, London, and Basic Books.

Little, I., and J.A. Mirrlees. 1990. Project appraisal and planning twenty years on. In Stanley Fischer, Dennis de Tray, and Shekhar Shah (eds.), *Proceedings of the World Bank Annual Conference on Development Economics*. Washington: The World Bank Publications Department.

Lowe, M., C. Papageorgiou, and F. Perez-Sebastian. 2016. The public and private MPK. Working Paper, IMF.

Lucas, R.E. 1988. On the mechanics of economic development. *Journal of Monetary Economics* 22: 3–42.

Lynde, C., and J. Richmond. 1993. Public capital and total factor productivity. *International Economic Review* 34: 401–414.

Mas, M., J. Maudos, F. Pérez, and E. Uriel. 1993. Competitividad, Productividad Industrial y Dotaciones de Capital Publico. *Papeles de Economia Espanola* 56: 144–160.

Munnell, A.H. 1990a. Why has productivity growth declined? Productivity and public investment. *New England Economic Review*. 2–22.

Munnell, A.H. 1990b. How does public infrastructure affect regional economic performance? *New England Economic Review*. 11–32.

Munnell, A.H. 1992. Policy watch: Infrastructure investment and economic growth. *Journal of Economic Perspectives* 6: 189–198.

Obstfeld, M., and K. Rogoff. 1996. *Foundations of international macroeconomics*. Cambridge, MA: MIT Press.

Otto, G.D., and G.M. Voss. 1994. Public capital and private sector productivity. *Economic Record* 70: 121–132.

Presbitero, A. 2016. Too much and too fast? *Journal of Development Economics* 120: 17–31.

Pritchett, L. 2000. The tyranny of concepts: CUDIE (cumulated, depreciated, investment effort) is not capital. *Journal of Economic Growth* 5: 361–384.

Romer, P.M. 1986. Increasing returns and long-run growth. *Journal of Political Economy* 94: 1002–1037.

Romp, W., and J. de Haan. 2007. Public capital and economic growth: A critical survey. *Perspektiven der Wirtschaftspolitik* 8: 1–140.

Sawyer, C.W. 2010. Institutional quality and economic growth in Latin America. *Global Economy Journal* 10 (4): 1–13.

Sturm, J.-E., G.H. Kuper, and J. De Haan. 1998. Modeling government investment and economic growth on a macro level: A review. In *Market behaviour and macroeconomic modeling*. London: MacMillan.

Sturm, J.-E., and J. De Haan. 1995. Is public expenditure really productive: New evidence for the USA and the Netherlands. *Economic Modeling* 12: 60–72.

Tanzi, V., and H.R. Davoodi. 2000. Corruption, growth, and public finances. IMF Working Paper 00/182.

Tatom, J.A. 1991. Public capital and private sector performance. *Federal Reserve Bank of St. Louis Review* 73: 3–15.

Warner, A.M. 2014. Public investment as an engine of growth. IMF Working Papers 14/148. International Monetary Fund.

World Bank. Independent Evaluation Group 2009. The World Bank's Country Policy and Institutional Assessment: An Evaluation.

World Economic Forum. 2016. The global competitiveness report 2016–2017.

Public Utility Pricing

Pierre B. Massé

Public utility goods and services are, for any given state of the economy and of technology, those to which the members of a society are regarded as entitled at reasonable charge, and which could not be satisfactorily distributed through the usual market channels.

The organizations responsible for distributing such goods and services, referred to here as public utilities, are characterized by the legal system which governs them, but not by the nature of the capital employed (whether public, private, or mixed).

The legal system governing public utilities is as a rule laid down under a grant of license and terms of reference. The granting authority is the central government, the local authority or some intermediate territorial entity. The grantee is bound by certain obligations, of which the main ones are the duty to ensure continuity of service and equality of treatment among users. It enjoys prerogatives of expropriation and easement which are especially valuable when setting up a *network* (whether railway, electricity, telephone, or drinking-water supply etc).

The ethical consideration that governs public utilities and is discussed in this entry is that of *benefit to the community.*

I. Pricing is a procedure involving the following preliminary stages: (a) a reasoned expectation of future demand, future technology and future costs of factors of production; (b) an optimal choice of technology and production factors so as to minimize the discounted total cost of the output that is in keeping at any time with demand, together with whatever prudent margin is considered necessary.

Pricing as such consists of calculating prices which maximize the benefit to the community, taking into account any economic, financial or social constraints. If there is neither rationing of the product, nor any unduly high margin of production potential, then the whole process is validated. Otherwise, it has to be adjusted in detail so as to remove any deficiencies or reduce any excess margin.

Historically, the problem of public utility pricing was first discussed with a modern outlook by Harold Hotelling in 1938. In France, after the work by Maurice Allais in 1943, an early reference text is the address by Gabriel Dessus of Electricité de France, 1949. This text uses the argument of the woodcutters' and miners' cases to illustrate the greater advantage – in terms of benefit to the community – of marginal-cost pricing compared with average-cost pricing. If it were possible to optimize continuously at every moment, then marginal cost would be defined without ambiguity. But because of the indivisibilities, as the same author showed using the paradox of the passenger for Calais, it is preferable to keep to long-run marginal cost (LRMC) and to smooth the irregularities of short-run marginal cost (SRMC).

The theory of LRMC which results from these analyses is developed in two monographs, Nelson (1964) and Morlat and Bessière (1971). Among these texts is an article by Marcel Boiteux (1956) on managing public monopolies which are under a duty to balance their budgets. This is a special constraint, but the method used is a general one and gives that paper value as a reference.

In striving to secure price-variations in keeping with economic efficiency, the public utilities were encouraged by contemporary advances in theoretical economics, initiated in the early 1950s by K.J. Arrow and G. Debreu (e.g. Debreu 1959). The worth of that approach is, in the first place, that it defines a commodity not just by its physical nature but also by the date and place of its supply, so justifying the varying of public utility prices. Next, it is valuable because it provides a modern and rigorous demonstration of the existence of a price equilibrium – combined with a Pareto optimum – based on the assumption under which economic transactors are compelled to optimize at *fixed prices* their behaviour as producers and consumers. A public utility that prices at LRMC is thus behaving 'appropriately' in its own field.

The dissemination and implementation of these principles encountered a number of difficulties. First, the theory led to price differences according to time of day and seasons, but the desired degree of variation could not be taken too far without making metering appliances complex and hence costly, as well as making for subtle differences in charging that might be ill-perceived by users.

More basic obstacles were the heritage from the past, the attachment to budgetary balance, a strong tendency to assign to a physically defined commodity or service (a cubic meter of water, a passenger-kilometre, a ton-kilometre or a kWh) a price that took no account of the time and place of its supply. The desire for equality made for a standardization of charges across the board. This is not to mention the political determination to redistribute incomes, which all too often took the convenient line of manipulating public utility prices. The economy may be 'imperfect' upstream of public utilities' inputs and downstream of their outputs. The Government's economic advisers must have a say regarding the impairments occasioned to economic efficiency in the name of social justice, in order to make the search for a compromise a consciously thought-out process.

II. The quarter-century after World War II was marked by a high level of employment and rapid growth. It was generally accepted in France that a moderate dose of inflation was 'the socially-acceptable price to pay' (Malinvaud 1983, p. 17). Optimizations at nominal fixed prices then had to be replaced by optimizations at *real* fixed prices.

International monetary instability, competition from Asia, and the oil shock in late 1973 resulted in an abrupt change in 1974 (see Dubois 1980). The world economy entered a zone of turbulence, and the West was struck by the scourge of unemployment, even though the social diffusion of almost thirty years of expansion staved off a recurrence of the tragedies of the Great Depression.

This situation led to the view in certain quarters that public utilities, quite apart from their duty to serve the public, should be required to fulfil an instrumental function of helping to restore the major equilibria of the economy (Courbis 1972). This concept gained further momentum from the post-war wave of nationalizations in Great Britain, France and Italy, followed by further nationalizations in France in 1981. The State, as the licensing authority, became the sole shareholder of a great many public utilities and hence, attempted to harness them to its general economic policy.

Experience has shown that the immediate response of many Western economies to a major and unexpected shock is equilibrium at fixed prices with under-employment (or other forms of rationing). What happens next depends upon the degree of flexibility of the economy. If sensitivity is shown to market indications, adjustment through prices can begin to take place. If, on the other hand, rigidity has its way, the managers of the economy are compelled to abandon the idea of a first-best optimum which would involve unacceptable real wages. They can then resort to macroeconomic models, either to compute shadow prices by second-best optimization (under pressure to reduce rationing), or to simulate the effects of changes in the behaviour of economic agents induced by the use of shadow prices supplied exogenously.

III. Long-run marginal cost is a conceptually desirable guide that is used empirically, and at times unconsciously, to an increasing extent; for example, the World Bank has endorsed its use for water and electricity supplies in developing countries (Munasinghe and Warford 1982). Even so, its application comes up against the heritage from the past, the temptation to unify prices for reasons of simplicity and/or egalitarianism, the concern to work towards greater social justice through non-apparent transfers of benefits, or again, the public utilities' own commercial considerations.

For these reasons, public utilities are at times compelled to compromise between practical or policy constraints and the purity of the theory. Knowledge of the theory nevertheless remains valuable to public utilities in democratic countries, since they can use it to gauge the extent of demands for transfers which sidestep parliamentary sovereignty.

Three Applications in France

The Railways

The railways, in France around the third decade of the nineteenth century, were very soon brought under the control of six private licensee companies with a tendency towards monopolistic practices. These companies were tied to the State by agreements, principally those of 1859 and 1921, under which a mutual financial support scheme was instituted, which took the practical form of a financial pool.

The Popular Front government of 1936, although it stopped short of outright nationalization, merged the networks in 1937 under a single company in which the State held a 51 per cent capital interest, the Société Nationale des Chemins de Fer Français (SNCF). After World War II, developments in the network were marked by two strategic changes: electrification using industrial current, with the backing of Louis Armand, and from 1980 onwards, the development of the new 'TGV' high-speed-train network which, from the passengers' point of view improved competitiveness on high-traffic trunk routes.

At the end of 1982, when the 1937 agreement expired, the assets of the SNCF reverted to the state under the terms of the agreement. Thereupon, a new outline act on domestic transport defined a 'new' SNCF with the status of an industrial and commercial public body, without resorting to the earlier principle of licensing.

Throughout this long process of change, passenger pricing is seen to have been a compromise between two tendencies: (a) the heritage from the past, which accounts in particular for nationwide standardization, a habit of thought deeply rooted in French mental attitudes; and (b) the pressures of economic efficiency in a climate of lively competition with the motor car, and then with airline flights.

As matters stand, the basic price per kilometre is the same everywhere, except on the new Paris–Southeast line, where the distances used in pricing are still based on the old routes, which are appreciably longer. A policy of variation according to time, gradually developed over the last 15 years, is based on supplements for certain trains, side by side with reductions in keeping with social as well as commercial considerations. A red, white and blue timetable, devised for the purpose, distinguishes off-peak days, normal weekends and some 30 or so major holiday departure dates. This policy is being pursued under new legislation and regulations which reserve to the State powers for approving pricing levels and structure.

For goods, railways were subject in the early days to two opposite influences. One was political and anti-monopolistic, and favoured standardizing charges; the other was economic, and motivated in particular by competition from road transport which 'creamed off' the more lucrative traffic. This led the railways to look for an efficient pricing structure (Hutter 1950). That took shape in particular through de-standardization adopted in 1961, on the basis of the Rueff–Armand report, and confirmed in 1967 after the Nora report: the terms of reference were relaxed, with greater flexibility for pricecompetitiveness. The legal provisions of 1982–3 seek to go still further by giving SNCF complete freedom in principle to set its own prices, while making it liable to observe 'the rules of fair competition'.

In practice, SNCF refrains from reducing its prices below LRMC – an important guide to its commercial policy. The intensity of competition with other modes of transport was considered sufficient to protect users against the possible 'abuse of a dominant position'.

Electricity

The licensed electricity distributing companies, which were formed in the late nineteenth century and grew up in the twentieth, were private-capital utilities. Their numbers tended to fall with takeovers, but even so in 1945 there were still over a thousand of them. Most were nationalized by an Act of 8 April 1946, under which Electricité de France (EDF) was formed as an industrial and commercial public body.

For EDF, forecasting demand, the choice of equipment and pricing are three links in the same chain. To begin with, demand was predicted according to the empirical rule, seductive in its simplicity, of a doubling in ten years. Very soon, however, less rudimentary models had to be developed, to take into account the general expansion of the economy and the prospects for the different types of use.

As a first stage, the optimum capital stock was sought by comparing, at the margin of the grid, the performance of the projected hydroelectric power stations with that of a reference fuel-burning power station. In the mid-1950s, EDF, with the help of computers which had just appeared on the market, conducted first linear, then non-linear programming for the entire grid, with the optimum solution generating dual prices that are associated with the easing of physical constraints.

All pricing policies are two-part (a fixed premium and proportional charges). The fixed-premium amount relates to demand-ratings subscribed by users in advance, and any excess demand drawn because too little demand was subscribed is billed separately. The charging scales comprise eight different periods of the year for major customers, and taking transactions' costs into account, the charging scheme becomes gradually simpler going down towards smaller-scale supply.

On account of the considerable impact of random events, control is necessary to balance supply and demand. Supply control (using hydraulic reservoirs and pumping stations) is supported by demand control, particularly through pricing. The requirements to subscribe to an amount of demand meets that objective by involving users in the coverage of risks. In addition, the rapid expansion of heating uses, which accentuates sensitivity to climatic hazards, has led the electricity suppliers to consider optional charging scales in which posted prices relate to periods of which customers know the duration (roughly 400 hours at the shortest and dearest), without knowing exactly when they will occur; this is currently specified by the supplier at the last moment.

There are few departures in France from the principle of LRMC pricing. The most notable was the practice of applying the same standard for charge to low-voltage sales in both urban and rural areas, when costs 40 years ago were appreciably lower in the former. However, the choices against efficiency that might have resulted from this practice remained limited, since low-voltage prices have virtually no impact on private individuals' choice of geographical location. What is more, cost differentials have narrowed appreciably with the expansion of consumption and increase in density of networks.

Concerning industry, public utilities in many countries are subject to – at times considerable – pressure to offer prices at less than cost to very large users, particularly in electrometallurgy and electrochemistry. As a rule, any really serious departures from the principles of pricing have been avoided. In the main, however, pricing remains a valuable guide for both customers – particularly industrialists – and producers in making their choices, including commercial policy choices.

Telecommunications

Ease of access to a telecommunications network at any point within national boundaries is a public utility for which pricing changes as the network expands and services develop (Hazlewood 1950; Squire 1973).

In the early stages, traffic was mainly local. A fixed charge reflecting the cost of average usage was the price of access to the service. Then the local networks spread, with small users alongside large. In pricing, transmission and switching costs were separated from access and rental charges, which are the responsibility of connection and management services.

In the stage which followed, trunk traffic services developed and, on account of their high value of usage, they subsidized for a time a portion of local traffic costs. A reduced rental charge was designed to draw in lower-income users. A further attraction to new subscribers was the consumption externality which is a particular feature of the telephone network: the utility of being connected to the network increases with its size, so that the increase in number of connections to the network stimulates demand and causes the service to spread by a snowball effect. This was observed in France in the 1970s, when the allocation of capital-investment priorities to telephone capital-equipment projects caused an explosion in demand which until then had been rationed by supply.

The network then entered its mature phase, with a slowing of growth in the size of the stock of mainstations and in traffic. The concern for economic efficiency led to a search for reductions in the cross-subsidizations induced by differentials between the prices charged and LRMC (Brunetière and Curien 1985), since trunk-call traffic was in fact subsidizing local traffic and connection costs, and calls were not charged at a fair rate taking account of time of day (Littlechild 1970) and duration. Recent measures taken in

France (Curien and Pautrat 1973) to make charges uniform according to distance when it exceeds 100 km (in 1969), to introduce a four-tier time-of-day variation (1984), and to charge local calls according to duration (1985) are tending to reduce cross-subsidizations internal to traffic. On the other hand, the high cost of the unit pulse and the lowering of access and rental charges are maintaining a high level of cross-subsidization of households by business. These cross-subsidies are in keeping with social concern for the more deprived groups. Existing subscribers also benefit from the external benefit of new subscribers coming onto the network, although the marginal investment costs of connection must remain profitable at the social rate of discount.

Changes in pricing are also made necessary by the diversification of uses. After access to the basic network, the routing of telephone traffic, the commodity supplied by the telecommunications service has become the bandwidth, a non-dedicated carrier that the user can use to transmit vocal information (telephone) or non-vocal information (data, text, pictures etc). Strict principles of economic rationality, that is, pricing the carrier regardless of content, will have to be relaxed in order not to deter people from using devices for which the potential market is promising. Finally, with the change in the regulatory environment and the emergence of deregulation, competition is appearing in the most profitable segments of the telecommunications market, prompting network managers to remove cross-subsidies.

See Also

- Communications
- Marginal and Average Cost Pricing
- Peak-load Pricing

Bibliography

Boiteux, M. 1956. Sur la gestion des monopoles publics astreints à l'équilibre budgétaire. *Econometrica* 24(1): 22–40.

Courbis, R. 1972. Tarifs publics et équilibre économique. *Economie et Statistique* 30: 19–27.

Curien, N., and C. Pautrat. 1973. An economic approach to telecommunications tariffs. *ITU Telecommunication Journal* April: 183–194.

de la Brunetière, J., and N. Curien. 1985. Les transferts de revenus induits par la tarification téléphonique entre catégories d'abonnés et entre types de prestations. *Annales des Télécommunications* 699.

Debreu, Gérard. 1959. *Theory of value: An axiomatic analysis of economic equilibrium*, Cowles Foundation monograph, No. 17. New York: Wiley.

Dessus, G. 1949. Document VI-5 of the UNIPEDE Congress in Brussels, May. Trans. in *International Economic Papers I*. London and New York: Macmillan.

Dubois, P. 1980. La rupture de 1974. *Economie et Statistique* 124: 3–20.

Hazlewood, A. 1950. Optimum pricing as applied to telephone service. *Review of Economic Studies* 18: 67–78.

Hotelling, H. 1938. The general welfare in relation to problems of taxation and of railways and utility rates. *Econometrica* 6(3): 242–269.

Hutter, R. 1950. La théorie économique et la gestion commerciale des chemins de fer: le problème tarifaire. *Revue Générale des Chemins de Fer* 69: 318–332.

Littlechild, S.C. 1970. Peak-load pricing of telephone calls. *Bell Journal of Economics and Management Science* 1(2): 191–210.

Malinvaud, E. 1983. *Essais sur la théorie du chômage*. Paris: Calmann-Lévy.

Morlat, G., and F. Bessière. 1971. *Vingt-cinq ans d'économie électrique*. Paris: Dunod.

Munasinghe, M., and J.J. Warford. 1982. *Electricity pricing. Theory and case studies*. Baltimore: Johns Hopkins University Press, for the World Bank.

Nelson, J.R. 1964. *Marginal cost pricing in practice*. Englewood Cliffs: Prentice-Hall.

Squire, L. 1973. Some aspects of optimal pricing for telecommunications. *Bell Journal of Economics and Management Sciences* 4(27): 515–525.

Public Utility Pricing and Finance

Frank A Wolak

Abstract

The theory of public utility pricing provides clear recommendations when the regulator and utility have same information about the underlying economic environment – the structure of demand and the production process. In reality, the utility has private information about the underlying economic environment, and the

incentives created by the regulatory process can cause it to exploit this information by producing in an inefficient manner. This insight complicates virtually all aspects of the theory of public utility pricing, and has led to theoretical characterizations of the public utility price-setting process as the solution to a mechanism design problem.

Keywords

Asymmetrical information; Averch–Johnson effect; Cost functions; Cost-of-service regulation; Increasing returns to scale; Inverse elasticity pricing rule; Multi-part tariffs; Multi-product firms; Natural monopoly; Optimal pricing; Price cap regulation; Principal and agent; Privatization; Public utility pricing and finance; Ramsey pricing; Regulatory contracts; Two-part tariffs

JEL Classifications

H4

Public utilities typically provide goods and services using a physical or virtual network infrastructure under a legal monopoly status. Public utilities can be privately owned, government-owned and customer-owned. Products provided by public utilities include electricity, natural gas, water, sewage treatment, waste disposal, public transport, telecommunications, cable television and postal delivery services. In the United States, all the different ownership forms can exist within the same industry. For example, in the electricity supply industry, there are privately owned, investor-owned and municipally owned utilities, and cooperative utilities owned by their customers.

Many explanations have been offered for the public utility industry structure. The standard economic efficiency argument is that the industry is natural monopoly, meaning that a single cost-minimizing firm is the least-cost way to serve the current level of demand. However, this logic relies on the implicit assumption that the single firm will produce in a cost-minimizing manner, which is unlikely to occur under government ownership or government regulation, for the reasons discussed below. In addition, although the current level of demand may be served at least cost by one cost-minimizing firm, this is unlikely to be that case for all future levels of demand as the number of customers or their purchasing power grows. Recognizing that public safety and health concerns argue for universal access to many of these services and the fact that the demand is very inelastic with respect to its own price leads to political economy explanations for this public utility industry structure. As Waterson (1988) notes, a government-owned or -regulated monopoly may better ensure that all customers have access to these services at reasonable prices.

Over the 100 years or more of state and federal regulation of public utilities in the United States there has been debate over what constitutes a reasonable price for goods and services of public utilities. A price that recovers the firm’s operating costs including return on its capital stock is generally considered to meet the legal standard of a reasonable price. This form of price regulation in the United States is often referred to as ‘cost-of-service’ regulation. However, as Joskow (1974) has persuasively demonstrated, the price-setting process for privately owned utilities in the United States does not guarantee the firm a fixed rate of return on its capital stock or full operating cost recovery. In that sense, to call this regulatory price-setting process ‘cost-of-service’ regulation is a misnomer. Joskow (1974, p. 325) states:

‘The rate of return aspect of regulation is merely a method by which a regulatory commission justifies its approval of price increases or major changes in rate structures. Without such triggering mechanisms the rate of return constraint is essentially inoperative.’ When the cost-of-service regulatory process operates it sets a price that allows the public utility an opportunity to recover its operating costs and the regulated rate of return on its capital stock through prudent operation.

If the firm earns a higher rate of return at this price because of superior management, then it is allowed to keep the revenue. If the firm earns a lower rate of return because of poor management, then shareholders must accept a lower rate of return. Only when the regulatory commission has overwhelming evidence that the higher or lower rate of return is due to extraordinary events beyond the control of the firm and not anticipated at the time the regulatory commission set the price will it make *ex post* adjustments to alter the public utility's regulated rate of return. A well-known example of an extraordinary event is a price change for fossil fuels used to produce electricity. The extreme volatility in oil, natural gas and coal prices since 1977 has led many regulatory commissions in the United States to implement fuel price adjustment clauses that automatically pass through in any input fuel price changes in the price of electricity. Baron and De Bondt (1979) discuss the impact of these fuel adjustment mechanisms on the investment and operating decisions of regulated electricity and natural gas utilities.

The terms and conditions surrounding this promise of full cost recovery through prudent operation is often referred to as the 'regulatory contract' between the regulatory commission and the public utility. This implicit contract requires the utility to serve all demand at the price set by the commission in exchange for a price that allows the utility the opportunity to recover its operating costs and a reasonable return to its capital stock. A major challenge to this regulatory contract is determining when imprudent operation is the cause of a failure to achieve full cost recovery.

For a number of reasons, unexpected events outside the control of the regulatory commission or utility and *ex post* opportunism by the regulator are often very difficult to distinguish from valid reasons for the regulatory commission to disallow price increases. Utilities typically require substantial investments in a network infrastructure that has limited alternative uses. The future demand for the public utility's services is uncertain, so there is a substantial risk that investments in network infrastructure will not be needed to serve the demand that exists when the investment is completed.

Several aspects of the regulatory process in the United States are designed to address the problem of the regulator setting a price that is insufficient to provide a reasonable return on past investments. The concept of a ratebase and the requirement that the regulatory commission sets a price that recovers operating costs and a reasonable return on the entire ratebase limits opportunistic behaviour on the part of the regulatory commission. To a first approximation, the ratebase is the sum of all past investments judged as prudent and therefore worthy of cost recovery by the regulatory commission. Phillips (1993, ch. 8) provides a detailed discussion of this concept. The requirement that the entire ratebase earns the regulated rate of return ensures that the current regulatory commission compensates the utility for investments that previous commissions have deemed prudent.

Gilbert and Newbery (1994) construct a dynamic model of the regulatory price-setting process where the commitment to allow the firm to earn a reasonable rate of return on a ratebase composed of past prudent investments results in a socially efficient level of investment by the regulated firm. Lyon and Mayo (2005) investigate the empirical relevance of regulators' opportunistic behaviour by examining the investment behaviour of regulated electric utilities and the propensity of the relevant state regulatory commissions to disallow investments by these utilities from entering the ratebase. Lyon and Mayo (2005) find little evidence that these cost allowances by the state regulatory commissions were due to opportunistic behaviour, and instead argue they were motivated by a desire to punish poorly managed firms.

Optimal Pricing of Public Utility Services with Full Information

Prices that adhere to the implicit regulatory contract of allowing full cost recovery only impose one restriction on the set of possible prices. For the case of a single-product utility that must set the same price for all customers, this restriction implies that the regulated price is equal to average total cost. However, virtually no public utilities sell a single product or are required to set a single price for all customers, so that regulatory commissions are free to pursue additional goals, besides the promise of cost recovery, in setting regulated prices.

This section discusses methods for setting economically efficient prices – those that maximize some social welfare function – under the simplifying assumption that the utility and the regulatory commission have the same amount of information about the utility's production process and demand. The remainder of this section assumes symmetric information between the regulated utility and the regulatory commission, so the commission can credibly set a price that only recovers the firm's minimum cost of serving its demand. Although the assumption of symmetric information about the production process and nature of demand between the utility and regulatory commission is unrealistic, the literature on optimal pricing for public utilities described in this section relies on this assumption.

Two-part tariffs relax the assumption that a single uniform price is charged to all customers for each unit of output. If the production of the good or service is subject to increasing returns to scale, setting price equal to the marginal cost of the last unit sold violates the legal requirement that the firm has an opportunity to recover total production costs. Coase (1946) addresses this problem by considering a regulated public utility producing a homogenous product with a monthly fixed cost of production, *F*, and a constant marginal cost, *c*. Coase (1946) argues that the total surplus maximizing two-part tariff sets the price of each unit consumed, *p*, equal to *c* and the fixed charge for each customer equal to *F*/*N*, where *N* is the number of customers served by the public utility.

If consumers differ in their willingness to pay for the product, then the surplus accruing to some consumers can be increased by the commission setting multi-part tariffs that charge different marginal prices for different ranges of monthly consumption. If the level of the monthly fixed charge necessary to recover total monthly fixed costs causes some consumers not to purchase the product, then a multi-part tariff can increase total consumer surplus. Assuming the marginal cost of a minute of telephone service is two cents per minute, setting a low monthly fixed charge and charging two cents per minute for the first 200 minutes of phone calls in the month and four cents per minute for all minutes above 200 minutes per month can allow the phone company to increase the number of consumers that benefit from having a telephone service without violating the promise of cost recovery. In this way, those consumers with the highest willingness will select through their consumption choice the higher marginal price, while those with the lowest willingness will select the lower marginal price, and virtually all consumers will pay a monthly fixed charge that does not cause them to disconnect from the telephone network. Brown and Sibley (1986, ch. 4) discuss consumer and producer welfare properties of multi-part tariffs.

The nature of the goods and services sold by public utilities often allows them to segment customers and to charge different prices for the same product. In addition, virtually all public utilities are multi-product firms, which implies that the regulatory process involves setting prices for all goods sold by the firm. Both of these circumstances provide opportunities for regulatory commissions to pursue objectives beyond the promise of cost recovery.

Consider the case of a homogenous product with increasing returns to scale in production that is sold to M different sets of consumers and a regulatory commission that can set a single price for each set of customers. Deriving the total surplus maximizing prices for all customer types subject to the constraint on cost recovery fits into the framework considered by Ramsey (1927). Let $CS_i(p_i)$ equal the consumer surplus accruing to consumers of type *i* and when they face price p_i, and $PS_i(p_i)$ equal the producer surplus from serving consumers of type *i*. Ramsey prices maximize the objective function $TS = \sum_{i=1}^{M} [CS_i(p_i) + PS_i(p_i)]$ subject to the cost recovery constraint, $F \leq \sum_{i=1}^{M} PS_i(p_i)$, where *F* is the firm's fixed cost. Let *c* equal the marginal cost of production and $\varepsilon_i(p_i)$ the own-price elasticity of the demand by customers of type *i* at price p_i. The solution to this constrained optimization problem yields the inverse elasticity pricing rule:

$$\frac{(p_i^* - c)}{p_i^*} = -\frac{k}{\varepsilon_i(p_i^*)}, i = 1, 2, \ldots, M$$

where k is some positive constant and the p_i^*, $i = 1, 2, \ldots, M$, are called Ramsey prices. These prices raise the revenue necessary to achieve full cost recovery with the smallest total surplus loss. Those consumer types with relatively more inelastic demands for the goods pay higher markups above marginal cost than other consumer types.

This same Ramsey-pricing logic can be applied to the case of multi-product public utilities. However, the simplicity of inverse elasticity pricing rule is complicated by the fact that customers can substitute across the products that the public utility offers. For example, in the pricing of postal delivery services, a business mailer has the option to use different US postal service products to communicate with its customers. To set Ramsey prices, the regulatory commission must know the multi-product cost function for postal products and the consumers' surplus associated with each postal product, which depends on the price charged for other postal products that the consumer can use as a substitute. As shown in Brown and Sibley (1986, ch. 3), both own-price and cross-price elasticities of demand now determine the total surplus maximizing markups that solve the Ramsey-pricing problem.

The properties of the regulatory pricing mechanisms described in this section rely on the assumption that the public utility's cost function is the minimum cost way to produce each vector of outputs. Specifically, let T equal the public utility's technology set, the pairs of vectors of input quantities, x, and output quantities, q, such that q can be produced using x. If w is the vector of input prices, then the minimum cost function is equal to

$$C(q) = \min_{x \geq 0} w'x \quad \text{subject to} \quad (x, q) \in T.$$

Although a price-taking profit-maximizing firm would like to produce along its minimum cost function, the structure of the regulatory process could cause a privately owned profit-maximizing regulated utility not to produce along its minimum cost function. Averch and Johnson (1962) present an example of a regulatory mechanism that causes a profit-maximizing firm not to produce in a least-cost manner. This work led to a massive theoretical and empirical literature exploring the distortions from minimum cost behaviour caused by regulatory price-setting processes.

State-owned utilities are likely to have even less incentive to produce in a least-cost manner. Besides the incentives provided by the regulatory price-setting process, earning revenues in excess of operating costs is just one of the many objectives that managers of state-owned companies must balance. As Waterson (1988, ch. 4) notes, state-owned utilities are often asked to pursue political or social goals that conflict with maximizing profits and therefore minimizing production costs.

Economists have begun to recognize the distinction between a public utility's observed cost function and its minimum cost function. The observed cost function, $CO(q)$, gives the firm's actual cost of producing output vector q given the incentives provided by the regulatory process. For example, in the case of a state-owned utility, political constraints could require a firm's management to hire a certain number of workers for each level of output despite the fact that it is technologically feasible to produce using fewer workers at each output level. In general, the value of the firm's observed cost function is greater than the value of its minimum cost function for the same level of output, and this difference can be substantial.

There is a vast empirical literature documenting violations of the assumption of cost-minimizing behaviour by public utilities. Christensen and Greene (1976) is a well-known example in the electricity supply industry and Evans and Heckman (1984) is one for the telecommunications industry. There are few empirical studies of regulated utility behaviour where the assumption of cost-minimizing behaviour is not rejected.

Optimal Pricing of Public Utility Services with Asymmetric Information

Public utility regulation has long been recognized as an example of agency relationship where the regulator (the principal) attempts to provide incentives for the public utility (the agent) to serve its demand in a least-cost manner at a price that only recovers observed costs. This link between the utility's production costs and the price it is allowed to charge creates the opportunity for the public utility to exploit its superior information about the production process and the demand it faces. This recognition has led researchers and regulatory policymakers to consider ways to either break this link between the price charged and the firm's observed cost or to design incentive mechanisms that exploit the link between the regulated price and the firm's observed cost.

Price-cap regulation attempts to break this link by committing to change the price the utility is allowed to charge according to a formula that cannot be altered for a sustained period of time. For a profit-maximizing firm, the price-cap mechanism creates the same incentive to minimize cost as the assumption of price-taking behaviour. In the United Kingdom (UK) this form of regulation is also called RPI minus X price-cap regulation because the rate at which prices are allowed to change on an annual basis is equal to rate of change in the retail price index (RPI) minus an X-factor chosen by the regulatory commission. Armstrong et al. (1994, ch. 3) describe the details of choosing an X-factor for a specific firm and extensions of this basic regulatory framework.

The major challenge associated with price-cap regulation is balancing the desire to commit to a pre-specified pattern for the X-factors for as long as possible against the fact that the longer duration of the commitment to this pattern of X-factors the greater the likelihood that the commitment will run afoul of the regulatory contract or political concerns. The experience of many public utilities in the UK privatized during the 1990s provides an instructive example of this phenomenon. According to Armstrong et al. (1994), the regulator for the water industry and the regulator for the electricity distribution sector initially committed to a five-year initial duration for values of the X-factors. However, in both these industries the regulator was forced to abandon these commitments before the end of the five-year period because of what was argued to by the water utilities to be inadequate revenues and what was argued by many consumers as excessive revenues in the case of the electricity distribution sector. Wolak (1998) discusses practical problems associated with implementing price-cap mechanisms and why they often evolve into an extremely ineffective form of cost-of-service regulation.

The development of game theoretical models of private information environments has led economists to derive optimal prices in this 'second-best' world. Baron and Myerson (1982) derive the total surplus maximizing prices when the public utility has private information about its production process not observable by the regulatory commission. The Baron–Myerson model assumes that the commission offers the firm a menu of prices and fixed fees that depend on the public utility's report of its private information. This report determines what price and fixed fee the firm is able to charge to its customers and therefore the revenues it is allowed to earn.

Recognition that informational asymmetries between the regulatory commission and utility can lead to significant distortions from minimum cost production and regulated prices that recover revenues substantially higher than these minimum costs has led to an explosion of theoretical work on the design of optimal regulated prices in this private information environment. Laffont and Tirole (1993) provide a comprehensive presentation of this literature.

Wolak (1994) attempts to quantify the cost of this private information in the regulatory process. Using a sample of California water distribution utilities, he estimates a full or symmetric information solution to the utility and regulatory commission interaction and a private or asymmetric information model of this interaction based on the Baron–Myerson (1982) model. Wolak (1994) finds that the asymmetric information model of the regulatory process provides a superior fit to observed data relative to the symmetric information model in addition to computing various estimates of the cost of the informational asymmetry between the firm and regulatory commission.

Concluding Comments

The explicit recognition of the impact of the private information possessed by the firm on the price set by the regulatory commission has increased the realism of the assumptions underlying models of the public utility price-setting process. Unfortunately, the form of the optimal price-setting mechanism derived from these models depends crucially on the source of informational asymmetries between the firm and the regulatory commission, as well as many other details of the economic environment. This implies that much more empirical work on actual regulatory price-setting processes is needed to implement these theoretical advances in actual regulatory practice.

See Also

- ▶ Averch–Johnson Effect
- ▶ Mechanism Design
- ▶ Principal and Agent (i)
- ▶ Principal and Agent (ii)
- ▶ Price Discrimination (Theory)
- ▶ Ramsey Pricing

Bibliography

Armstrong, M., S. Cowan, and J. Vickers. 1994. *Regulatory reform: Economic analysis of the British experience*. Cambridge: MIT Press.

Averch, H., and L. Johnson. 1962. Behavior of the firm under regulatory constraint. *American Economic Review* 52: 1052–1069.

Baron, D.P., and R.R. De Bondt. 1979. Fuel adjustment mechanisms and economic efficiency. *The Journal of Industrial Economics* 27: 243–261.

Baron, D.P., and R. Myerson. 1982. Regulating a monopolist with unknown costs. *Econometrica* 50: 911–930.

Brown, S.J., and D.S. Sibley. 1986. *The theory of public utility pricing*. Cambridge: Cambridge University Press.

Christensen, L., and W.H. Greene. 1976. Economies of scale in U.S. electric power generation. *Journal of Political Economy* 84: 655–676.

Coase, R.H. 1946. The marginal cost controversy. *Economia* 13: 169–189.

Evans, D., and J. Heckman. 1984. A test of subadditivity of the cost function with an application to the Bell system. *American Economic Review* 74: 615–623.

Gilbert, R.J., and D.M. Newbery. 1994. The dynamic efficiency of regulatory constitutions. *RAND Journal of Economics* 25: 538–554.

Joskow, P. 1973. Pricing decisions of regulated firms: A behavioral approach. *Bell Journal of Economics and Management Science* 4: 118–140.

Joskow, P. 1974. Inflation and environmental concern: Structural change in the public utility price regulation. *Journal of Law and Economics* 17: 291–327.

Laffont, J.-J., and J. Tirole. 1993. *A theory of incentives in procurement and regulation*. Cambridge: MIT Press.

Lyon, T.P., and J.W. Mayo. 2005. Regulatory opportunism and investment behavior: Evidence from the U.S. electric utility industry. *RAND Journal of Economics* 36: 628–644.

Phillips, C.F. 1993. *The regulation of public utilities*. Arlington: Public Utilities Reports.

Ramsey, F.P. 1927. A contribution to the theory of taxation. *Economic Journal* 37: 47–61.

Waterson, M. 1988. *Regulation of the firm and natural monopoly*. Cambridge: Basil Blackwell.

Wolak, F.A. 1994. An econometric analysis of the asymmetric information regulator-utility interaction. *Annales d'Economie et de Statistique* 34: 13–69.

Wolak, F.A. 1998. Price cap regulation in newly privatized industries. *Oxford Energy Forum* 34: 12–14.

Public Works

P. Bridel

Keywords

Balanced budget multiplier; Beveridge, W. H; Crowding out; Effective demand; General gluts; Hawtrey, R.G.; Hoarding; Kahn, R. F.; Keynesianism; Multiplier analysis; Neoclassical synthesis; New classical macroeconomics; Pigou, A. C.; Primary and secondary labour markets; Propensity to consume; Public works; Ricardo, D.; Saving equals investment; Treasury view; Underconsumptionism; Unemployment; Webb, B. and S.

JEL Classifications

B0

Public works to relieve unemployment is an idea as old as the pyramids. From the cathedrals of the

Middle Ages to the dams of the Tennessee Valley Authority, recent centuries abound with examples of such government-supported schemes. However, it is only fairly late in the history of economic theory that economists started devoting some analytical thought to this problem. In fact, it was around the turn of the 20th century, with the appearance on the Continent of the first systematic work on business cycles, that economists stopped looking at unemployment as a 'pathological' problem to be tackled primarily as one of charity and relief. Even down to the British 1909 *Report of the Poor Law Commission* a majority still considered unemployment as no challenge to economic theory: if properly applied, in the long run, the normal laws of value and distribution would see to a solution to this problem.

A notable exception to this general neglect is, of course, the heated discussions that took place between early 19th-century classical economists during the so-called 'general glut controversy'. In the course of that well-known debate on the self-adjusting capacities of the economic system, public works as a means to combat unemployment came many times to the forefront of the argument. Ricardo, James Mill and Say fairly easily won the day against public works. However, though dissenters like Malthus, Lauderdale, Bentham and Sismondi rejected Ricardo's doctrine on the stability of the economic system and Say's rigmarole about the equality of supply and demand, they involved themselves in a logically very damaging acceptance of the dominant Turgot–Smith saving-is-investment doctrine, together with the aggravating assumption that hoarding cannot take place. Trapped in such contradictory statements, these dissenting authors never managed to put their case convincingly either against Ricardo's self-adjustment doctrine or, of course, in favour of contra-cyclical public works. Showing very little interest in 'immediate and temporary effects [and fixing his] whole attention on the permanent state of things' (1952, vol. 7, p. 120), Ricardo was left in a very strong theoretical position to dismiss arguments in favour of credit expansion and/or public works to reduce unemployment. Furthermore, his lack of a theory of output (that is, his full-capacity utilization assumption) allowed Ricardo to close in an even tighter way a line of argument that was to dominate economic theory for nearly a century, and economic policy for an even longer period. Anticipating nearly word for word Churchill's 1929 Budget Speech and what was to be known in the 1920s and 1930s as the 'Treasury View', in 1821 Ricardo already observed in the Commons:

> When [I] heard the honourable members talk of employing capital in the formation of roads and canals, they appeared to overlook the fact that the capital thus employed must be withdrawn from some other quarter. (1952, vol. 5, p. 32)

This 'Ricardian view' on public investment was to be perpetuated down the 19th century; after having been defended and illustrated by most classical economists, and notably by Mill, it even survived the marginalist revolution: Marshall never advanced much beyond it and Böhm-Bawerk subscribed substantially to its conclusions.

However, during the 1880s, Foxwell and Hobson were among the first to take up the challenge again. More preoccupied with the social effects of the 'irregularity of employment' linked with price fluctuations, than with strictly theoretical questions, they both called for abandoning the charity and relief approach and for turning the fight against unemployment into a major objective of economic policy. Foxwell concludes his analysis by favouring (unlike most leading economists) a slightly rising price level 'for more regular employment'. Hobson, for his part, suggests solutions to prevent under-consumption crises at the root of unemployment by a redistribution of income to encourage 'high consumption'. Despite a few attempts by local authorities to provide relief work to the unemployed, nothing systematic surfaced in Britain in the realm of economic policy despite growing concern about unemployment.

In the history of public works doctrine, the *Minority Report of the Poor Law Commission* (*R*oyal Commission on the Poor Law and the Unemployed 1909) clearly marks a watershed: for the first time ever the minority commissioners advocated a systematic counter-cyclical policy of public works and investment to smooth out cyclical fluctuations and to stabilize employment and the level of economic activity. However, this fundamental turnabout was not confined to Sidney and Beatrice Webb, largely responsible (with A.L. Bowley for the statistical material) for drafting the *Minority Report*. In the same year as the *Report*, Beveridge in *Unemployment, a Problem of Industry* broadly supported its main conclusion, and a few years later the Webbs in their volume *The Prevention of Destitution* (1911) offered a more elaborate version of their original argument.

However, the first modern refutation of the 'Ricardian view' by a professional economist is due to Pigou in his 1908 inaugural lecture as Marshall's successor at Cambridge. Worked out again in his *Wealth and Welfare* (1912) and *Unemployment* (1913), Pigou's argument became the standard pre-Keynesian case for public works: without having to resort to the notion of budget deficit, Pigou is able to demonstrate that public spending can increase aggregate employment and does not simply divert it from the private to the public sector:

> . . . it is probable that only part of the extra taxes people pay would be taken from funds they would otherwise have devoted at that time directly or indirectly to wage payment. Hence, the true result of relief works and so on is not to leave the aggregate amount of unemployment in the country unaltered, but to diminish that amount. (1908, pp. 27–8)

In other words, and in the modern parlance of the balanced budget multiplier, the taxpayer's marginal propensity to consume is smaller than 1, while the government's marginal propensity to consume is unity: the net effect of such a tax increase is clearly expansionary. Pigou's failure to assess his argument quantitatively (in a multiplier-like fashion), his scepticism about the degree of labour mobility between the private and the public sectors and the resulting weak impact of his argument on policymakers do not detract however from his originality. Public works, even without any budget deficit, can lessen unemployment.

This became a standard argument for most economists, well before the First World War. Robertson gave his 'cordial support' to Pigou's analysis and, for the first time brought forward in his *Industrial Fluctuation* the symmetrical idea that, through public works, governments would 'in times of depression [use] savings [that] are *not* otherwise so applied' (1915, p. 253).

With the notable exception of Hawtrey in the 1920s, most British economists followed Pigou's lead even if they were bitterly arguing about the best way to close the gap between saving and investment. Hawtrey for his part never departed from his early critical position against the *Minority Report* outlined in *Good and Bad Trade* (1913, p. 260). He reiterated many times, most notably in his 1925 article, that the public works doctrine 'overlook[s] the fact that the Government by the very fact of borrowing for its expenditure is withdrawing from the investment market savings which would otherwise be applied to the creation of capital' (1925, p. 104). As an adviser to the Chancellor of the Exchequer, and despite the broad professional consensus in favour of public works, Hawtrey won the day in the British Treasury. Under successive Conservative and Labour Governments, the 'Treasury View' remained official wisdom for years (for a detailed discussion of that evolution see Hutchison 1953, pp. 409–23 and Winch 1969, pp. 94–113). Similarly, during the 1932 US presidential election Roosevelt kept attacking the outgoing Hoover administration for not balancing the budget. It is only with the rearmament in Germany and with the Second World War in other Western countries that the 'Treasury View' all but disappeared from the politicians' standard set of arguments.

In his famous 1931 multiplier article, Kahn managed once and for all to dispose of this 'Treasury View'. Kahn's intentions were, however, empirical, not theoretical. He set out not so much to point out the then generally accepted argument that an increase in government investment would generate 'secondary employments' but to provide 'a stronger case' for public works than that which 'had always been recognized' by giving a precise estimate of this multiplier effect. The ratio of secondary employment to primary employment was thus given its first formal expression (1931, p. 12)

However, and to dispel a very common confusion, with a multiplier equation strictly in the

Pigovian tradition, Kahn did not anticipate the theoretical core of Keynes's *General Theory,* that is, the principle of effective demand. In Kahn's model, even if additional public spending has a multiplier effect on output and employment, such public works do not affect the discrepancy between saving and investment: as for Pigou and Robertson, previously, 'unused' savings are simply brought back into circulation. In short, for Kahn 'the whole point of a policy of public works is that it enables an increase in the rate of home investment . . . without that *fall* in the rate of interest that would be necessary if we were relying on private enterprise' (1931, p. 26).

In the *General Theory,* with his principle of effective demand, Keynes added a new theoretical dimension to Kahn's multiplier approach to the public works doctrine. An increase in public investment not only leads to an increase in employment and output but *generates* an excess of income over that required for consumption (via a marginal propensity to consume smaller than 1) so that the volume of savings will increase until saving once again equals investment. Furthermore, it is clear that since saving and investment are brought into equality by variations in the level of income, and not by changes in interest rate, there need not be full employment. In Keynes's theoretical framework, public works and government investment are thus no longer seen as temporary remedies to cyclical fluctuations in private investment, but as a necessary component of an aggregate demand the deficiencies of which are no longer automatically cared for, even in the long run, by an interest-rate mechanism.

This proposition (and the income-adjustment principle underlying it) could of course find no place in the traditional analytical framework. However, systematic short-term counter-cyclical public works policies were soon to be integrated by mainstream economists within the so-called neoclassical synthesis and played in most countries a major role in post-war economic policies. Victim of its success during the 1950s and 1960s, this Keynesian (as opposed to Keynes's) public works doctrine ran into serious practical and theoretical problems in the early 1970s. The growing size of public sectors, the inflationary wave resulting largely from an excessive use of expansionary fiscal policies and a rapidly growing theoretical dissatisfaction of the economic profession with the neoclassical synthesis brought back the 'Treasury View' onto the theoretical agenda. However even if with the successive advent of the monetarist school and the 'governmental impotence' theorem of the New Classical School, the 'crowding out' hypothesis is currently enjoying a new lease of life, few economists and even fewer politicians would dispute today the importance of modulating government spending in the course of the cycle.

See Also

▶ Hawtrey, Ralph George (1879–1975)

Bibliography

Beveridge, W. 1909. *Unemployment: A problem of industry.* London: Longmans, Green.

Hawtrey, R.G. 1913. *Good and bad trade*. London: Constable.

Hawtrey, R.G. 1925. Public expenditure and the demand for labour. *Economica* 5: 38–48. Repr. in R.G. Hawtrey. *Trade and credit.* London: Longmans, 1928.

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

Kahn, R.F. 1931. The relation of home investment to unemployment. *Economic Journal* 41: 173–198. Repr. in R.F. Kahn. *Selected essays on employment and growth.* Cambridge: Cambridge University Press, 1972.

Pigou, A.C. 1908. *Economic science in relation to practice. Inaugural lecture given at Cambridge*. Cambridge: Cambridge University Press.

Pigou, A.C. 1912. *Wealth and welfare*. London: Macmillan.

Pigou, A.C. 1913. *Unemployment*. London: Williams & Norgate.

Ricardo, D. 1951–73. *The works and correspondence of David Ricardo.* 11 vols, ed. P. Sraffa. Cambridge: Cambridge University Press.

Robertson, D.H. 1915. *A study of industrial fluctuation.* London: LSE Reprint, 1948 (with a new Introduction).

Royal Commission on the Poor Law and the Unemployed. 1909. *The minority report of the poor law commission.* 2 vols. London: Fabian Society.

Webb, S., and B. Webb. 1911. *The prevention of destitution*. London: Longmans, Green.

Winch, D. 1969. *Economics and policy. A historical study.* London: Hodder & Stoughton.

Pufendorf, Samuel von (1632–1694)

Arild Sæther

Abstract

Pufendorf studied in Leipzig and Jena. His first work, *Elementorum Jurisprudentiae Universalis* (1660), earned him a professorship at Heidelberg. In 1668 he moved to Lund. His works *De Naturae et Gentium* (1672) and *De Officio Hominis et Civis* (1673) were translated, spread all over Europe, and entered the curricula at most Protestant universities. Pufendorf's natural law writings include ethics, jurisprudence, government and political economy. A society in which individuals exchange to satisfy their needs brings with it growth, commerce, markets, prices and money. This theory laid the foundation for the progress of economics as a science.

Keywords

Commercial society; Demand and supply; Externalities; Grotius, H.; Hobbes, T.; Hutcheson, F; International law; Locke, J; Money; Natural law; Price; Property; Pufendorf, S. von; Self-interest; Smith, A; Snob effect; Taxation; Veblen effect; Voting

JEL Classifications

B31

Pufendorf was born in Dorfchemitz, Saxony, Germany on 8 January 1632. He matriculated at Leipzig University in 1650. First he studied theology, but found it dogmatic and turned to philosophy, philology and history. After two years he moved to Jena, concentrating on mathematics, the Cartesian demonstrative method and the natural law writings of Grotius and Hobbes.

After completing his Magister degree in 1658 he secured an engagement as a tutor in the family of the Swedish ambassador in Copenhagen. Shortly thereafter hostilities broke out between the two Nordic rivals. Disregarding diplomatic privileges, the Danes seized the Swedish retinue and accused Pufendorf of espionage. During eight months of harsh captivity, with no access to learned books, he reflected on his university studies and wrote down a system of jurisprudence. After his liberation, he journeyed with his pupils to the Netherlands, where his work was published in 1660 as *Elementorum Jurisprudentiae Universalis* (Elements of Universal Jurisprudence). This work is considered the first useful textbook on natural law and it earned Pufendorf an enviable reputation and a professorship at the University of Heidelberg.

Here he published, anonymously, his historical and political work *De Statu Imperii Germanici* (On the Constitution of the German Empire). It contains a devastating criticism of the condition of public law within the Empire resulting from the Thirty Years' War, and suggested a path to its regeneration through a European commonwealth of sovereign states based on natural and international law. The imperial censor banned the book, but it was reprinted time and again, translated into several languages, and distributed across Europe. By 1710, some 300,000 copies had been printed in Germany alone. Pufendorf's reputation was now extended to nonacademic circles. He achieved both fame and criticism.

In 1668 Pufendorf moved to the newly established university in Lund, Sweden. In 1672 he published his magnum opus *De Naturae et Gentium* (On the Law of Nature and Nations) in eight books, and the year after an abridged popularized version, *De Officio Hominis et Civis* (On the Duty of Man and Citizen), in two books.

In all, 44 editions of his major work have been published. It has been translated into English, French, German and Italian. His popularized version became, in modern parlance, an international bestseller. It has been translated into nine European languages, published in more than 150 editions in tens of thousands of copies. For more than 100 years they were among the most read academic books. The classicists Locke, Montesquieu, Rousseau, Hutcheson, Hume, Smith and countless more all studied and built on his works. Due to Pufendorf's works and reputation, natural law became part of university studies in jurisprudence, philosophy and ethics at most Protestant universities.

A new war resulted in the closing of Lund University in 1677. Pufendorf became royal historiographer in Stockholm. In the following years, he introduced empirical studies of the archives, and published 33 volumes of historical studies. He is regarded as a progenitor of 19th-century historicism.

In 1688 Pufendorf moved to Berlin as historiographer and judicial councillor at the court of Prussia. He continued his works on historical and theological issues. He died on 16 October 1694 of blood poisoning, contracted on a return journey from Stockholm, where he had been elevated to the nobility as a baron. He is buried in St. Nikolai Church in Berlin.

Pufendorf attempted in his works on natural law to mediate and unify Hobbes's natural law doctrine of 'egoism' and 'a war of all against all' with Grotius's natural law doctrine of 'man's inclination towards society'. His writings include ethics, jurisprudence, government, and political economics. These are seen as integral parts of a totality.

The foundation for his treatment is his theory of human behaviour, where the driving force is the interaction between man's self-interest and his existence as a social being. Man seeks society with his fellow man for the fulfilment of his own needs and desires. Man's sociable inclination is not innate; it must be cultivated. He also used his theory of the social man to create his historical account of the rise of property when society changes from hunting and gathering through agriculture to a commercial society. Individuals in a commercial society will need goods and services produced by others, because their own time and resources will fail to give them many necessary goods. On the other hand, individual men can contribute many things to the use of others. However, this will come to nothing if these individuals could not exchange and barter their different goods and services. When a society based on private property grew, it therefore brought with it commerce, the growth of markets, the creation of prices and the introduction of money. The theoretical foundation of a commercial society, in which all individuals attempt to satisfy their own needs and thereby satisfy the need of others, is therefore the cornerstone in his natural law theory.

Price is divided into *ordinary and eminent*. The former is found in the properties of goods and services in so far as they afford service and pleasure for man. The latter is found in money as a common standard for their measurement. The price of a good or service is determined by the interaction between 'the aptitude' (utility) of it and the scarcity of it: in modern parlance, demand and supply. The price will rise towards a level where it covers the normal costs that accrue during production and transport. Lack of need (demand) lowers the price, but price will also be lowered if the number of suppliers increases. Pufendorf therefore comes very close to a Marshallian demand-and-supply analysis. In addition, the price will change if the quantity of money changes. Pufendorf seems to recognise the Snob and Veblen effects, externalities and differences in the elasticity of demand.

Pufendorf also presents his views concerning the state and the distribution of power, the state's right to tax, and the principles of taxation. Here he discusses weighted voting, qualified majorities and what has been known as single-peaked preferences.

It was the diffusion of these theories through popularization which laid the foundation for the progress of economics as a science.

See Also

- ▶ Hutcheson, Francis (1694–1746)
- ▶ Smith, Adam (1723–1790)

Selected Works

1660. *Elementorum Jurisprudentiæ Universalis Libri Duo* [The elements of universal jurisprudence]. Oxford: Clarendon Press, 1931.

1672. *De Jure Naturae et Gentium Libri Octo* [On the law of nature and nations].

New York/London: Oceana Publications Inc./ Wiley & Sons Ltd., 1933; reprinted 1964.

1673. *De Officio Hominis et Civis*. Trans. M. Silverthorne as *On the duty of man and citizen according to natural law*, ed. J. Tully. Cambridge: Cambridge University Press, 1991.

Bibliography

Gaertner, W. 2005. *De Jure Naturae et Gentium*: Samuel von Pufendorf's contribution to social choice theory and economics. *Social Choice and Welfare* 25: 231–241.

Hont, I. 1986. The language of sociability and commerce: Samuel Pufendorf and the theoretical foundations of the 'four-stages theory'. In *The languages of political economy in early modern Europe*, ed. A. Pagden. Cambridge: Cambridge University Press.

Hutchison, T. 1988. *Before Adam Smith: The emergence of political economy 1662–1776*. Oxford: Basil Blackwell.

Luig, K. 1972. *Zur Verbreitung des Naturrechts in Europa. Tijdschrift voor Rechtsgeschiedenis*, vol. 40. Groningen: Wolters-Noordhoff N.V.

Othmer, S.C. 1970. *Berlin und die Verbreitung des Naturrechts in Europa*. Berlin: Walter de Gruyter & Co.

Sæther, A. 1996. Pufendorf: The grandfather of modern economics. In *Samuel Pufendorf und die europäische Frühaufklärung*, ed. F. Palladini and G. Hartung. Berlin: Akademie Verlag.

Sæther, A. 2000. Self-interest as an acceptable mode of human behaviour. In *The Canon in the history of economics: Critical essays*, ed. M. Psalidopoulos. London/New York: Routledge.

Pump Priming

Leon H. Keyserling

The expression 'pump priming' gained nationwide vogue during the Roosevelt New Deal 1933–9. It referred to US Government spending accompanied by deficit financing to promote economic recovery from the Great Depression, which peaked in 1933 at an unemployment rate of 24.9 per cent and GNP about 30 per cent below 1929 in real terms. The expression suited F.D.R.'s political skills and innate economic and financial conservatism because so many people then used hand pumps and their experience was that pouring in a little bit of water for a short time started a copious and sustained flow in normal fashion. When the US started to become 'the arsenal of democracy' in 1940 and later on entered World War II, the use to the term 'pump priming' ended.

Pump priming cannot significantly be discussed academically as 'theory', but only by specifying what was done and how it worked. Despite factors which limited both size and pace measured against the catastrophic business and human conditions spawned by the Great Crash, the efforts were varied and enterprising; they bespoke the 'experimentalism' and 'try anything' mood of the President and his advisers.

For all practical purposes, the job-oriented programmes started early in 1933 and peaked within about a year. The Civilian Conservation Corps was empowered to employ less than 250 thousand people at $30 a month relief-level wages. The Federal Emergency Relief Administration (FERA) was armed with $500 million for grants in aid to the states, but the required matching grants from the states was a severe limitation, and the means test was applied to ultimate beneficiaries. The National Industrial Recovery Act included $3.3 billion for a Public Works Administration (PWA). The Civil Works Administration (CWA), unlike FERA, could operate works projects directly. This was the swiftest job programme; by January 1934, more than four million jobs resulted, but most CWA funds were drawn from PWA, and thus did not add to net outlays. By early 1934, opposition to CWA placed it within the FERA, where it could finance only state and local activities, help only those previously on relief, and pay only subsistence rather than prevailing wages.

PWA pump priming under Secretary of the Interior Harold L. Ickes dealt with construction, repair, and improvement of highways, public building and any other publicly owned facilities; conservation; low cost housing and slum clearance; and 'any other project of a character previously constructed or carried on by public authority or with public aid to serve the interest of the general public': and loans could be made even to some private corporations. The CWA under Harry

Hopkins dealt, *inter alia*, with roads, airports, school facilities and personnel, playgrounds and swimming pools, sewers, and combating insect pests. The jobs included writers and painters.

The potency of these job programmes must be judged by the gap between the actual size of the 1933 economy and one operating at reasonably full resource use. Measured in uniform dollars, actual GNP in 1933 would have had to be more than 43 per cent higher to close the gap between it and the 1929 performance. But this vastly understates the problem. For without growth to absorb a growing potential in labour force, technology, and know-how, an economy languishes. To be as close to full utilization as the economy was in 1929, GNP in 1933 would have had to be almost 72 per cent higher than it actually was. Looking at the GNP averaged during 1933–7, the gap measured in these two ways was 16 per cent and almost 51 per cent, respectively. And looking at the GNP averaged during 1933–9, the gap was more than 10 per cent and almost 49 per cent, respectively.

Contrasted with these criteria of recovery needs – even if modified considerably – Federal outlays and deficits were minimal, to state it charitably. Outlays were about $4.6 billion in 1933, or less than 8.3 per cent of GNP, and the deficit of $2.6 billion was less than 4.7 per cent of GNP. In 1939, outlays of about $8.8 billion were less than 9.7 per cent of GNP, and the deficit of $3.9 billion was less than 4.3 per cent of it. And only about two-thirds of these outlays, classified as 'extraordinary', were for pump priming *and* for the costs of new 'reform' measures; about one-third, for defence and customary domestic, were classified as 'ordinary'. All in all, increased Federal spending through pump priming outlays was much less than the decline (due to the Great Depression) of state and local outlays for similar purposes.

Despite this spotty and far from adequate performance, the repeated assertion that the New Deal and especially pump priming 'failed' because there were more than 8 million unemployed in 1940 even after the start of the 'arsenal of democracy' is not supportable. Comparing 1937 and 1933, unemployment dropped from 12.830 million to 7.700 million, or 40 per cent (annual average 12.0 per cent). The drop in the rate of unemployment was from 24.9 per cent to 14.3 per cent, or a 42.6 per cent drop (annual average 13.0 per cent). The GNP real growth rate was 45.6 per cent (annual average 9.8 per cent).

The President and the Congress were strongly inclined to reduce pump priming when substantial albeit dismally insufficient recovery occurred. The cutbacks intensified greatly by 1937, and this brought on for two years one of the sharpest economic turndowns on record. Even so, unemployment at 9.480 million in 1939 was 26.1 per cent below 1933 (annual average reduction 4.9 per cent). The unemployment rate fell to 17.2 per cent, a reduction of 30.9 per cent (annual average 6.0 per cent). By 1940, due partly to increased military spending, unemployment was down to 8.12 million. During no periods of comparable length after World War II, despite smaller difficulties and larger know-how, were the rates of reduction in unemployment and real GNP growth nearly as large as during the pump priming era.

'Reform' programmes not called 'pump priming' also aided recovery: bank deposit insurance, regulation of the stock market and other financial practices, farm price supports, stimulation of home building, credit at lower interest rates, and some other measures. Because these 'reform' measures were as much the product of expanded government responsibility as pump priming, pouring public money into the economy and entering into the deficit Budgets, they are not really separable from appraisal of the new public philosophy which pump priming most clearly embodied.

Most important in the long run: although the humanistic new public philosophy, with pump priming setting the tone, emerged because of national disaster, it evoked an *enduring* people's awareness, sensed by political leaders, that *their* Government should exercise immensely enlarged responsibility for economic performance and human well-being; that huge public spending is not anathema; and that balancing the economy should take precedence over balancing the Federal Budget.

World War II did not submerge this heritage. Tremendous war expenditures, claiming almost half of GNP, were financed 50 per cent by deficits, and primed the pump as never before. Thus, during the war years, the average annual real economic growth rate was about 9 per cent, and the unemployment rate dropped to less than one per cent. And because ‘equality of sacrifice’ reflected the social concern of 1933–9 pump priming, living standards rose more rapidly than ever before.

Since the war, national policies and their results have oscillated, but from a plateau of economic and social responsibility far higher than before the pump priming precept and example. Public action to help the unemployed has become a ‘given’; reform programmes have been built upon and new ones added; Federal *domestic* spending during recent years has ranged around 17 per cent of GNP, compared with only 5.5–6.5 per cent during prewar pump priming. Consequently, although there have been frequent recessions, there have been no depressionary downturns which came every seven years or so even before 1929.

However, the most recent years, under both Democratic and Republican Administrations, have brought considerable retrogression. ‘Big Government’ has come to be excoriated and Federal support of vital domestic programme slashed; Federal Budget domestic spending in ratio to GNP has fallen from 17.50 per cent in fiscal 1985 to 15.89 per cent in the President’s 1986 Budget; and pronounced efforts to balance the Budget at the expense of the economy have in fact generated low GNP growth and unparalleled deficits. The goal of full employment, although on the statute books, has in practice been completely surrendered.

To be sure, New Deal pump priming, because it started with 24.9 per cent unemployment, did not within seven years come close to unemployment rates as low as the highest during any year after World War II which provided a postwar starting point of only one per cent unemployment. But the policies and programmes during the pump priming era 1933–9 pressed for and achieved very substantial *reduction* in unemployment, contrasted with the complacency during the mid-1980s despite a rate two-and-a-third times as high as the 2.9 virtually full employment rate in 1953. From 1977 to the second quarter of 1985 (similar in length to 1933–40), the unemployment rate *increased* from 7.1 per cent to 7.3 per cent, and unemployment rose from 7.4 million to 8.4 million. This brought the disgraceful and menacing rate of unemployment among black teenagers up to 39.2 per cent, immensely higher than the overall unemployment rate of 24.9 per cent in 1933 and 17.2 per cent in 1939. And compared with the high real growth rate in GNP during the pump priming era, the real growth rate, punctuated by four recessions, averaged annually less than 3 per cent during the period, and only about 2 per cent from 1979 forward.

The compelling lesson of the more than half a century reviewed is this: we should not underestimate the pump priming era. From it we can learn much and benefit greatly.

See Also

- ▶ Budgetary Policy
- ▶ Built-in Stabilizers
- ▶ Public Works

Purchasing Power Parity

Lucio Sarno

Abstract

This article expounds the purchasing power parity (PPP) hypothesis as a theory of exchange rate determination. The long history of PPP and its contribution to thinking in international finance is discussed, with reference to implications both for open economy theory and for economic policy. The large literature on empirical testing of the validity of PPP is reviewed, with particular reference to work carried out since 1990.

Keywords

Aggregate demand; Balassa, B.; Beggar-thy-neighbour; Bretton Woods system; Cassel, G.; Co-integration; Common factors; Consumer price indices; Crawling peg exchange-rate regime; de Malynes, G.; Deflation; Exchange rate depreciation; Exchange rate dynamics; Exchange rate regimes (empirics); External competitiveness; Fixed exchange rates; Flexible exchange rates; Gold standard; Gross domestic price deflators; Haberler, G.; Harrod, R.; Hyperinflation; Imperfect competition; Indexation; Inflation; Inflationary expectations; Information costs; International arbitrage; International capital flows; International finance; International portfolio diversification; International trade barriers; Johnson, H.; Keynes, J. M.; Kravis, I.; Labour market contracts; Law of one price; Microfoundations; Monetary policy; Monte Carlo methods; Mundell, R.; Mundell–Fleming model; Nominal exchange rates; Nominal interest rates; Nurkse, R.; Open economy macroeconomics; Price stability; Productivity differentials; Purchasing power disparities; Purchasing power parity (PPP); Quantity theory of money; Rational expectations; Real exchange rates; Real interest rates; Real price ratios; Ricardo, D.; Rueff, J.; Salamanca School; Samuelson, P.; Spatial arbitrage; Spatial price differentiation; Sticky prices; Strong (absolute) purchasing power parity; Terms of trade; Trade barriers; Unit roots; Viner, J.; Weak (relative) purchasing power parity; White noise; Wholesale price indices; Yeager, L

JEL Classifications

F3

Purchasing power parity (PPP) is a theory of exchange rate determination. It asserts (in the most common form) that the exchange rate change between two currencies over any period of time is determined by the change in the two countries' relative price levels. Because the theory singles out price level changes as the overriding determinant of exchange rate movements, it has also been called the 'inflation theory of exchange rates'.

The PPP theory of exchange rates has somewhat the same status in the history of economic thought and in economic policy as the quantity theory of money: by different authors and at different points in time it has been considered an identity, a truism, an empirical regularity or a grossly misleading simplification. The theory remains controversial, as does the quantity theory of money, because strict versions are demonstrably wrong while soft versions deprive it of any useful content. In between there is room for theory and empirical evidence to specify the circumstances under which and the extent to which PPP provides a useful, though not exact, description of exchange rate behaviour.

The analogy with the quantity theory of money holds particularly in the effects of monetary disturbances. The latter theory fails to hold exactly when disturbances are primarily monetary, for instance in the course of hyperinflations, because changes in the expected rate of inflation generate systematic movements in velocity that break the one-to-one link between money and prices. In the same way, monetary disturbances cause exchange rate movements that at least temporarily deviate from PPP, implying changes in the exchange rate-adjusted relative price levels or 'real' exchange rates. It is true that when the economy, following a major monetary disturbance, has settled down again the cumulative changes in money, prices and the exchange rate will tend to be close to each other. In that sense PPP holds. The same is decidedly not true, however, in the course of the disturbance.

And in the long run, just as changes in real income or financial innovation bring about trend changes in velocity that destroy the one-to-one relationship between the money supply and prices, there also are trend deviations from PPP: productivity growth differentials between countries, for example, lead to trend changes in real exchange rates.

Statement of the Theory

Let *pi* and p_i^* represent the price of the *i*th commodity at home and abroad, stated in home and foreign currency respectively, and *e* the exchange rate. The exchange rate is quoted as the number of units of domestic currency per unit of foreign money. Further, let *P* and *P** be the price level at home and abroad quoted in the respective currencies.

The strong or absolute version of PPP relies on the 'law of one price' in an integrated, competitive market. If we abstract from all frictions, the price of a given good will be the same in all locations when quoted in the same currency, say dollars: $p_1 = ep_i^*$. Consider now a domestic price index $P = f$ $(p_1, \ldots, p_i, ..p_n)$ and a foreign price index $P^* = g$ $\left(p_1^*, \ldots, p_i^*, \ldots, p_n^*\right)$. If the prices of each good, in dollars, are equalized across countries, and, if the same goods enter each country's market basket with the same weights (that is, the homogeneous-of-degree-one $g(\cdot)$ and $f(\cdot)$ functions are the same), then by definition *absolute* PPP prevails. The law of one price in this special case extends not only to individual goods but also to aggregate price levels. Spatial arbitrage then takes the form of the *strong* (or absolute) version of PPP:

$$e = P/P^* = \frac{\$\,\text{price of a standard market basket of goods}}{\pounds\ \text{price of the same standard basket}} \tag{1}$$

where the RHS is the common multiple of the price of each good in one currency and in the other. Specifically, if $p_i/p_i^* = k$ for all *I*, we then have $e = P/P^* = k$. Note now the implication of absolute PPP. Whatever the monetary or real disturbances in the economy, because of instantaneous, costless arbitrage the prices of a common market basket of goods in the two countries, measured in a common currency, will be the same or $P/eP^* = 1$ at all times.

There can be no objection to (1) as a theoretical statement. Objections arise, however, when it is interpreted as an empirical proposition. In fact the (spot) prices of a given commodity will not necessarily be equal in different locations at a given time. Transport costs and other obstacles to trade, especially tariffs and quotas, do exist and hence the location of delivery does matter. Therefore we would not expect the price even of an ounce of gold of a specified fineness always to be the same in New York and in Calcutta. The fact that prices of the perfectly homogeneous commodity are not equalized across space at every point in time does not suggest market failure; it may simply reflect the inability to shift commodities costlessly and instantaneously from one location to the other. Information costs and impediments to trade stand in the way of strictest spatial equalization of price. But these impediments to trade do not preclude the possibility that common currency prices of any given good in different locations should be closely related and, indeed, arbitraged. They just will not be literally equalized. Impediments to trade and imperfection of competition, of course, also make possible spatial price differentiation, thus further limiting strong PPP.

The *weak* (or relative) version of PPP therefore restates the theory in terms of changes in *relative* price levels and the exchange rate: $e = \theta P/P^*$ where θ is a constant reflecting the given obstacles to trade. Given these obstacles an increase in the home price level relative to that abroad implies an equiproportionate depreciation of the home currency:

$$\widehat{e} = \widehat{P} - \widehat{P}^* \tag{2}$$

where $\hat{a}$ denotes a percentage change.

Equation 2 is the statement of PPP as it was applied by Gustav Cassel to an analysis of exchange rate changes during the First World War.

> The general inflation which has taken place during the war has lowered this purchasing power in all countries, though in a different degree, and the rates of exchange should accordingly be expected to deviate from their old parities in proportion to the inflation of each country. At every moment the real parity is represented by this quotient between the purchasing power of the money in the one country and the other. I propose to call this parity '*purchasing power parity*'. As long as anything like free

> movement of merchandise and a somewhat comprehensive trade between the two countries takes place, the actual rate of exchange cannot deviate very much from this purchasing power parity. (Cassel 1918, p. 413)

Absolute PPP in (1) was stated in terms of the relative prices in different currencies and locations of a *given* and common basket of identical goods. Going from there to relative PPP as in (2) may merely be a way of circumventing the qualifications arising from transport costs or obstacles to trade. But often more is involved because the shift, in practice, leads to a use of PPP in terms of particular price indices such as consumer price indices (CPIs), wholesale price indices (WPIs), or gross domestic price (GDP) deflators. Once that is done we go beyond the law of one price because the shares of various goods in the different national indices may not be the same and the goods that enter the respective indices may not be strictly identical as is clearly the case for nontradable goods.

Once shares in the indices differ or commodities are not strictly identical, the appeal to the law of one price can no longer serve as support for PPP. Now PPP can hold, even in the weak form, only if the disturbances satisfy the conditions of the homogeneity postulate of monetary theory. The homogeneity postulate asserts that a purely monetary disturbance, with all equilibrium relative prices left unchanged, will lead to an equiproportionate change in money and all prices, including the price of foreign exchange. In this very special experiment PPP holds even if the law of one price does not apply. The constancy of real variables under the assumption of a purely monetary disturbance (that is, an unanticipated, nonrecurrent increase in money) assures that once the economy has adjusted the exchange rate depreciation matches the inflation of any individual price or the price of any market basket so that (2) applies. To appreciate the difference of this experiment from absolute PPP, note that under these conditions (2) could even be stated in terms of indices of nontradable goods prices.

PPP theory as a theory of equilibrium must be supplemented by an adjustment mechanism. In the case of identical commodities the theory is simply that of spatial arbitrage. But when the goods are not strictly identical more is required. A high degree of substitution in world trade is generally assumed to be the mechanism through which exchange rate-adjusted prices are kept in line internationally. A further point concerns causation. In much of the literature, especially in the writings of Cassel, exchange rates adjust to prices. But there is an important alternative tradition that singles out exchange rate depreciation as an independent source of inflation.

Criticism of PPP focuses on systematic ways in which relative price changes destroy the strict validity of PPP. Keynes, although strongly supporting the idea of PPP as a broad guide, recognized these possible departures from purely monetary disturbances:

> If on the other hand these assumptions are not fulfilled and changes are taking place in the 'equation of exchange', as economists call it, between the services and products of one country and those of another, either on account of movements of capital, or reparation payments, or changes in the relative efficiency of labour, or changes in the urgency of the world's demand for that country's special products, or the like, then the equilibrium point between purchasing power parity and the rate of exchange may be modified permanently. (Keynes 1923, p. 80)

This limitation of PPP led Samuelson to argue: 'Unless very sophisticated indeed, PPP is a misleading, pretentious doctrine, promising what is rare in economics, detailed numerical prediction' (Samuelson 1964, p. 153).

History

Versions of the PPP theory have been traced to the Salamanca School in sixteenth century Spain and to the writings of Gerard de Malynes appearing in 1601 in England. The Swedish, French and English bullionists in the second part of the eighteenth century and in the early nineteenth century present further statements of PPP. Particularly noteworthy is the Bullion Report in England:

> Whether this 13$\frac{1}{2}$per cent, which stands against this country by the present exchange on Lisbon, is a real difference of exchange, occasioned by the course of trade and by the remittances to Portugal on account

> of government, or a nominal and apparent exchange occasioned by something in the state of our currency, or is partly real and partly nominal, may perhaps be determined by what your committee have yet to state. (Great Britain 1810, p. ccxxii)

During the nineteenth century classical economists, including in particular Ricardo, Mill, Goschen and Marshall, endorsed and developed more or less qualified PPP views. This history is reviewed and discussed in Viner (1937), Schumpeter (1954), Holmes (1967), and Officer (1984).

Even though PPP theory was well established by the time of the First World War, the forceful use and development of the theory by the Swedish economist Gustav Cassel has made him the outstanding protagonist of the theory. He turned the theory into a paradigm, with all the necessary trappings: an alleged challenge to gold standard orthodoxy, a catchy name, a formula, and the claim of empirical support for the new view.

Cassel's first contributions to the subject were published in 1916 in the *Economic Journal*. He presented the inflation theory of exchange rates and proceeded to a demonstration using price level and exchange rate data for the belligerent countries, the United States, and Sweden. J.M. Keynes as the editor appended a footnote drawing attention to the contribution and noting his surprise that, war disturbances notwithstanding, PPP should hold. A further challenge was the implication of PPP that the pre-war par with gold might not be re-established or, more guardedly, might require a powerful deflation in a country like Britain.

Cassel never abandoned an uncompromising PPP view of exchange rates even though he already in 1918 started recognizing the possibility that exchange rates might transitorily diverge from PPP. A decade later he made a clear statement of his final position:

> The fact that the rate of exchange corresponding to Purchasing Power Parity possesses such a remarkable stability is a sufficient reason for regarding Purchasing Power Parity as the fundamental factor determining the rate of exchange and for classifying all other factors that may influence the rate and perhaps make it deviate from the Purchasing Power Parity as factors of secondary importance, most suitably grouped under the head of 'disturbances'. (Cassel 1928a, p. 16)

He identified three groups of disturbances: actual and expected inflation or deflation, new hindrances to international trade, and shifts in international movements of capital. Although these disturbances are recognized, their quantitative effect on deviations from PPP is invariably seen as 'confined within rather narrow limits' (Cassel 1928a, pp. 28–9). In insisting on the proposition that deviations from PPP are limited and transitory, Cassel neglected to pay close attention to the determinants of purchasing power disparities. Even though he recognized that inflation first leads to undervaluation, and stabilization leads later to an overvaluation (Cassel 1928b, p. 26), he never took these ideas further. His emphasis was on PPP. But he pointed out with some merit (Cassel 1928b) that, without some quantifiable concept of PPP, a sensible discussion of over and undervaluation could hardly begin.

Keynes (1923, ch. 3) took up PPP, crediting Ricardo with the invention and Cassel with the name. Keynes recognized PPP as an important empirical possibility. Giving it all the right qualifications he still endorsed it for all practical purposes:

> This theory does not provide a simple or ready-made measure of the 'true' value of the exchanges. When it is restricted to foreign-trade goods, it is little better than a truism. When it is not so restricted, the conception of purchasing power parity becomes much more interesting, but is no longer an accurate forecaster of the course of the foreign exchanges. Thus defined 'purchasing power parity' deserves attention, even though it is not always an accurate forecaster of the foreign exchanges. The practical importance of our qualifications must not be exaggerated. (Keynes 1923, pp. 77–8)

Cassel received support for PPP from the monetary disturbances of the 1913–1928 period. Extensive PPP studies were conducted for the US government (see Young 1925) and for the League of Nations. PPP emerged in the discussion of the resumption of the pre-war gold par in Britain in 1925, and Jacques Rueff used wage-based PPP to calculate an appropriate par for France's stabilization under Poincaré in 1926–1928. But while it became a regular tool of applied macroeconomics, there was also plenty of controversy. Viner (1937) challenged the doctrinal view that classical economists had a concept of PPP, arguing that without the notion of a price level PPP could not be conceived. In fact Viner had little patience with PPP. The opposition is easily recognized today: Viner and other critics always reacted to the overstated claim that PPP must hold as a matter of fact or of theory, pointing out that only a purely monetary disturbance provided the theoretical or practical experiment in which PPP would apply. For them PPP as a theory was simply misstated and as a practical proposition overrated.

A new wave of interest in PPP emerged at the end of the Second World War, when once again exchange rates had to be set following the wartime suspension of trade and convertibility. Renewed interest in PPP followed in the late 1950s and early 1960s. Yeager (1958) and Haberler (1961) emphasized the practical usefulness of PPP and highlighted the role of high price elasticities in international trade as the factor supporting PPP. High elasticities in world trade would ensure that real disturbances had only small effects on relative prices, thus establishing more nearly the conditions under which exchange rate movements predominantly reflect differences in monetary experiences.

In the late 1930s Harrod had drawn attention to the fact that divergent international productivity levels could, via their effect on wages and home goods prices, lead to permanent deviations from Cassel's absolute version of PPP. This idea had already been developed by Ricardo and has now become central to work on international real income comparisons. Balassa (1964) and Samuelson (1964) elaborated similar ideas to argue that there are systematic trend deviations from PPP. This 'productivity bias' to PPP is discussed in more detail below.

PPP had yet another intellectual upturn with the move to flexible exchange rates in the early 1970s. The then fashionable 'monetary approach to the balance of payments' developed by Robert Mundell (1968, 1971), Harry Johnson and their students readily adapted to become a PPP-based monetary approach to the exchange rate (see Frenkel and Johnson 1975, 1978; Mussa 1979). The exchange rate under strict PPP conditions was interpreted as a monetary phenomenon. The absolute version of PPP in (1) above combined with the quantity theory of money for each country ($MV = PY$ and $M^*V^* = P^*Y^*$) yielded the key equation determining exchange rates by relative money supplies, velocities and real incomes:

$$e = (M/M^*)(V/V^*)(Y^*/Y). \qquad (3)$$

Empirical research on the 1920s and on the very early data of the 1970s initially seemed to lend support to PPP and the monetary approach.

But large movements in real exchange rates of the 1970s led to the currently dominant PPP scepticism. The new direction following the Mundell–Fleming model (Mundell 1968; Fleming 1962) of the 1960s emphasized fluctuations in real exchange rates or the terms of trade (import prices relative to export prices) arising from the discrepancies between flexible, forward-looking asset markets and asset prices, and short-run sticky prices and wages. Work on exchange rate dynamics (Dornbusch 1976) developed these ideas about transitory deviations from PPP in a rational expectations context.

Concern with PPP continued to be very active in the late 1970s and the early 1980s. The real exchange rates of the main currencies underwent large, persistent fluctuations with important effects on trade flows and resource allocation. At the same time currency experiments in Latin America involved dramatic real appreciations with ruinous consequences for several countries. Sometimes in history there was bafflement as to how, all things considered, PPP could work so closely. This time, however, the surprise was on the other side: how could real exchange rates get that far out of line? We now review in more detail the theory and the evidence for deviations from PPP.

Purchasing Power Disparities

Qualifications to PPP take one of several forms. Departures from PPP can be 'structural' in the sense that they arise systematically in response to new and lasting changes in equilibrium relative prices. Alternatively, they occur in a 'transitory' fashion as a result of disturbances to which the economy adjusts with differential speeds in goods and assets markets. These qualifications imply that even the weak or relative form of PPP cannot be expected to hold closely.

These disparities arise primarily for the following reasons. First, the terms of trade may change as a consequence of changes in trade patterns. Second, economic growth systematically affects the relative price of home and tradable goods. Third, monetary and exchange rate changes bring about transitory deviations in real price ratios and in PPP as a consequence of imperfectly flexible wages and prices.

Structural Departures

The literature is replete with qualifications to PPP singling out particular real disturbances that change equilibrium relative prices. Thus it has been recognized since Ricardo that real prices of home goods are high 'in countries where manufactures flourish'. It also has been argued that the 'price level is high in borrowing countries'. The Ricardo–Harrod–Balassa–Samuelson theory provides a framework for these ideas.

Consider a Ricardian model where the law of one price applies to tradable goods and where there is also a home good. With perfect competition and constant returns, prices are given by unit labour costs. We define as *R* the relative consumer price levels of two countries measured in a common currency:

$$R \equiv P/eP^*. \tag{4a}$$

With identical homothetic tastes and the law of one price the international component of price indices is the same in both countries and hence cancels out in (4a). The relative price level is then determined by the relative prices of home goods in the two countries, measured in a common currency. Let h and h^* be the levels of productivity in tradable goods (at the competitive margin) relative to home goods in each country. It is easily shown (see Dornbusch et al. 1977) that the relative price level then reduces to:

$$R = R(h/h^*), R' > 0. \tag{4b}$$

A uniform rise in tradable goods productivity at home would bring about a rise in the relative price level of the home country or a real appreciation. The mechanism is the following: with the law of one price applying to tradable goods, increased productivity in the tradable goods sector increases wages in that industry and hence raises economy-wide wages. But, without accompanying productivity gains in the home goods sector, costs and prices there must rise and hence the growing country's relative price level increases as shown by (4b).

In (4b) above the national productivity relatives h and h^* are measured in the tradable goods sector at the competitive margin. Shifts in technology, tastes, commercial policies or labour force growth will all change the equilibrium competitive margin and hence will change the real exchange rate. Thus real factors, as the literature since Ricardo has recognized, will introduce systematic departures from PPP. For example, a shift in world demand towards the home country's goods would raise the relative wage and reduce the range of goods produced by the home country. The rise in the relative wage, given productivity, raises the relative price level of the home country. Likewise an increase in spending relative to income (that is, borrowing or a current account deficit) will lead to a rise in the relative price level of the spending country.

A variant of the Ricardian productivity differential model as an explanation for the relatively low price of non-tradables in poor countries has been advanced by Kravis and Lipsey (1983) and Bhagwati (1984). They rely on differences in factor endowments and factor rewards rather than differences in production functions. In the poor labour-abundant country, the labour-using non-tradable services can be produced at a lower cost than in the rich, capital-abundant country. Whichever is the model, this effect, as we discuss below, has found ample support in empirical research on international real income and price comparisons.

Transitory Deviations

There is no difficulty in accepting that prices of close substitutes or even identical goods could diverge across space at any point in time. This would be the case because, in the shortest time period, transportation and information costs make arbitrage difficult or even impossible. These difficulties would explain that PPP holds up to a constant and white noise error (see Aizenman 1986). But in fact we have to explain relatively *persistent* and often *large* deviations from PPP. These can arise from divergent speeds of adjustment of the exchange rate compared with wages and prices. Particularly when flexible exchange rates behave like asset prices while wages are determined by long-term contracts, there is room for relative prices to show relatively persistent deviations from PPP.

Theoretical approaches to support the relative stickiness of prices can rely on the presence of long-term labour contracts combined with oligopolistic pricing in goods markets. A model of imperfect competition is essential because the less-than- perfect degree of substitution is a key ingredient in PPP deviations. Less-than-perfect substitution means that we are not dealing with the law of one price and arbitrage but with firms' decisions to set relative prices. A suggestive framework is the Dixit and Stiglitz (1977) model of product diversification with imperfect competition. Given constant returns and labour as the only factor each firm will set prices as a fixed and common markup over wages. In the world market for the products of a particular industry the relative price of domestic and foreign variants of the product is determined by relative unit labour costs measured in a common currency:

$$p/ep^* = w/w^*e \tag{5}$$

where w and w^* denote unit labour costs at home and abroad in the respective currencies. Given sluggish wages, for contract reasons or otherwise, exchange rate movements will be one-for-one reflected in changes in the real exchange rate.

The assumption that firms base their pricing entirely on home cost, as appears in this model, leaves no room for the alternative of spatial price differentiation. There is as yet no definitive or even large body of literature that develops the industrial organization aspects of pricing under flexible and volatile exchange rates (see Dornbusch 1987).

Early Empirical Evidence

There is little doubt that the prices of primary commodities traded on major organized exchanges in different locations are fully arbitraged when literally all adjustments for contracts (maturity, delivery terms and location, and so on) are made. But much available evidence suggests that PPP in the strong or weak version does not apply in the same fashion to manufactured goods. This lack of close conformity with PPP is as true for individual commodity prices as it is for aggregate price indices. Moreover, the absence of a very tight PPP relation appears to hold especially during major monetary dislocations.

Studies of high-inflation episodes appear to offer support for PPP in that they show close *cumulative* movements of internal prices and the exchange rate. But even here the evidence is deceptive, as becomes clear when one looks at relative prices, which do show large variations. Indeed, particularly during high inflation the differing frequencies of adjustments of wages, prices and the exchange rate introduce considerable variability in relative prices which disappears only in the most intense stages of hyperinflation where all pricing comes to be based on the exchange rate. Kravis and Lipsey (1978) and Isard (1977) have shown tests of the law of one price at the level of narrowly defined manufactured goods. These studies established that for the same good (or highly substitutable goods) there are quite definitely persistent price discrepancies between domestic and export prices, between domestic and import prices, and between export prices to different markets.

Empirical studies on time series PPP relationships for aggregate price indices since the mid-1960s also show evidence of persistent deviations. Once relative prices are not strictly constant PPP will perform differently depending on the particular price index chosen for comparison.

Commonly the choice is among CPIs, WPIs, and GDP deflators. WPIs are often ruled out on the argument that conceptually they are poorly defined, being neither producer nor consumer price indices. The preference is most often given to CPIs and GDP deflators, which have a clear methodological definition.

As a measure of the departure from PPP, research shows that, for the post Bretton Woods period since 1971 or so, bilateral comparisons of exchange-rate-adjusted inflation rates (that is, comparisons of inflation rates measured in a common currency) reveal that the correlation coefficients are much lower than unity, which is the theoretical value implied by the weak form of the PPP hypothesis.

The strong deviations from PPP can likewise be found by looking at relative prices, in which case one would compare the variability of relative price indices (the standard deviation expressed as a fraction of the mean), measured in a common currency and using the United States as the numeraire country. The data for these relative price variability measures show a large increase in variability in the shift from fixed (Bretton Woods period) to flexible (post-Bretton Woods) exchange rates, suggesting that real exchange rates are approximately as volatile as nominal exchange rates (Baxter and Stockman 1989).

The evidence on deviations from PPP leaves little doubt that they have been large and persistent. To pin down the major sources of these movements, however, is significantly more difficult. Among the chief explanations are capital flows induced by internationally divergent monetary–fiscal mixes interacting with sluggish wages and prices. Thus it would appear that a country that shifts in the direction of tight money and easy fiscal policy, for example, will experience real appreciation.

Besides these dominant macro-shocks there is, of course, a host of other factors. Jacob Frenkel has observed in this context:

> The experience during the 1970s illustrates the extent to which real shocks (oil embargo, supply shocks, commodity booms and shortages, shifts in the demand for money, differential productivity growth) result in systematic deviations from PPP ... It should be noted, however, that to some extent the overall poor performance of the purchasing power parities doctrine is specific to the 1970s. During the floating rate period of the 1920s, the doctrine seems to have been much more reliable. (Frenkel 1981, pp. 694–5)

The lack of solid empirical evidence in support of PPP extends to the assumption that divergent price developments ‘cause’ exchange rate depreciation. From the study of experiences of high inflation it is clear that in some instances capital flight and exchange depreciation precipitated increases in inflation. In fact Nurkse (1944) makes much of the point that expectations acting via capital flight on the exchange rate, and not actual money and prices, often initiate an inflationary episode.

With respect to structural PPP deviations, there is some evidence that establishes that over time real exchange rates, rather than showing constancy or a tendency to fluctuate around a constant level, in fact exhibit a distinct trend. Productivity levels or real incomes influence systematically the relative prices of tradable and non-tradable goods within a country and hence international relative price levels across countries and across time.

In the context of an international income comparison project, Kravis and associates have constructed indices of relative national price levels using an absolute price comparison approach. Drawing on a detailed sample of prices they construct matched sets of the price of individual commodity groups in a particular country relative to a reference country. For commodity i the relative price is p_i/p_i^*, where the $p's$ are measured in the respective countries’ currencies with an asterisk denoting the reference country. Using an arithmetic average with weights a_i, given by final expenditure shares, a PPP index is defined:

$$\mathrm{PPP} \equiv \sum a_i \left(p_i/p_i^* \right). \qquad (6)$$

The expenditure shares a_i used in the weighting may be those of either one of the countries or some other appropriate weighting scheme. The Kravis real price level of a country (relative to the reference country) is defined as the PPP index in (6) divided by the actual exchange rate:

$$\text{Kravis real price level} \equiv \text{PPP}/e. \tag{7}$$

This real price level definition represents a measure of the deviation from the law of one price at the aggregate level.

Kravis and Lipsey (1983, p. 21) report the results of a cross-section study of 34 countries where the 1975 real price level as defined in (7) of the sample of countries (relative to the United States) is explained by the countries real income compared with that of the United States. The evidence shows that the higher a country's relative income is, the higher is its relative price level. Work by Hsieh (1982) using a time series approach further supports the extensive evidence on divergent productivity trends as a source of structural PPP deviations. It must be noted, though, that the evidence on structural deviations continues to be challenged by Officer (1984).

Implications of Purchasing Power Disparities

The possibility that exchange rate movements do not conform to tight PPP patterns poses important issues for macroeconomic measurement, linkages, and policy. We review here several implications.

Real Income Comparisons

With strict PPP based on the law of one price, the purchasing power of a given income in one country and currency can be compared with the purchasing power of the income of any other country by simply measuring incomes in a common currency. If one income is 20 times larger than the other, measured in the same currency at actual exchange rates, then its command over goods and services is 20 times larger. But the fact that PPP does not hold leads to systematic biases in the comparisons. Specifically, as the work of Kravis and associates (1978, 1982, 1983) has shown, the real income of poor countries is severely underestimated when actual exchange rates are used to make the comparison. The low relative price of non-tradables in poor countries (due to the productivity differential discussed earlier) yields for poor countries true purchasing power of income significantly above what exchange rate-converted income suggests.

Note that the biases are particularly large for countries whose incomes are only a small fraction of the US levels, so that productivity differential effects play a maximal role. The poorer a country, the lower is the real price level. An interesting point is that these real price level differences apply both to goods and to services. One reason they also apply to goods is that these always have a local retail component which, on account of labour costs (though perhaps not transport), will tend to be low in poor countries. For low-income countries actual real income is two to three times what exchange rate-converted incomes suggest. These structural deviations from PPP, of course, would be invariant under a purely monetary disturbance so that the weak form of PPP still applies.

Interest Rate Linkages and PPP

Under perfect international mobility of capital and risk-neutral speculation there is a linkage between nominal interest rates and the anticipated rate of depreciation, which is given by the open economy Fisher equation:

$$i = i^* + x \tag{8a}$$

where i and i^* are the nominal interest rates at home and abroad and x is the anticipated rate of depreciation of the home currency. Adding and subtracting anticipated inflation rates on both sides yields an equation in terms of inflation-adjusted or real interest rates:

$$r^* = r - \dot{R}/R. \tag{8b}$$

Real interest parity, according to (8b), prevails when the real interest differential equals the expected rate of real appreciation, $\dot{R}/R$. From the real interest parity condition it is clear that under exact PPP the real exchange rate is constant. In the absence of restrictions on capital flows, real interest rates must therefore be strictly equalized across countries.

The real interest parity equation has two interesting implications. A first one is the linkages between the level of real exchange rates and monetary policy. Suppose that in a medium-term macroeconomic context, following a disturbance, the actual real exchange rate adjusts only gradually to the trend level R' according to the process: $R/R = (1/\lambda)$ $(R' - R)$. Here $1/\lambda$ is the speed of adjustment, which depends among other things on the extent to which wages and prices are sticky. Combining this process with (8b) yields an equation for the equilibrium real exchange rate:

$$R = R' + \lambda(r - r^*). \tag{9}$$

The result shown here is that, when real interest rates at home exceed those abroad, the real exchange rate will be low or appreciated relative to its trend value. A tightening of monetary policy, by raising real interest rates, would thus bring about a (transitory) real appreciation. Equation 9 emerges from the dynamic Mundell–Fleming models and is often thought to explain real exchange rate movements and their tendency to return only gradually to their long-run value.

A second way to look at (8b) draws on the fact that the tradable–non-tradable goods distinction has implications for real exchange rates. Suppose the law of one price holds for tradable goods and that shares in the two countries' price indices are the same. Then, as argued before, the real exchange rate is equal to the relative price of non-tradable goods (in a common currency) in the two countries. Structural disturbances such as differential productivity growth or changes in aggregate demand will now have a systematic impact on relative non-tradable goods prices and hence real interest-rate differentials. Specifically, the country with the higher growth rate of productivity has a rising relative price of home goods and thus has a lower real rate of interest. As another example consider a country where aggregate demand is transitorily high. The real price of home goods will be high, but falling. Accordingly, the real interest rate will be higher than that abroad. Deviations from PPP, trend or short-run, thus introduce an equilibrium international interest-rate differential.

PPP deviations affect interest differentials another way. In (8a) above we assumed risk neutrality. But, once risk-averse speculators are admitted, the possibility that exchange rate movements could deviate from a strict PPP pattern introduces portfolio risk associated with the currency composition of the portfolio. PPP deviations are thus one basic motive for international portfolio diversification. A risk premium will appear and among the determinants of this premium is the variability of the real exchange rate. The risk premium will be an increasing function of real exchange rate uncertainty.

Exchange Rate Policy

In Cassel's view even small deviations from PPP would bring about large changes in trade flows and hence a rapid discipline to move prices back into line internationally. But the reversion towards PPP has often not been quick and deviations from PPP have taken more nearly the pattern of persistent swings in a country's external competitiveness. The changes in competitiveness in turn have implied large swings in external balances, in output and in employment in the tradable goods sector. Changes in exchange rates that deviate from PPP at the same time influence the path of a country's inflation: real depreciation increases inflation and real appreciation dampens inflation. These effects of purchasing power disparities make the exchange rate an important issue in macroeconomic policy.

Countries with high inflation cannot afford a fixed exchange rate since the loss in external competitiveness would soon lead to excessive and growing external deficits and high unemployment. If freely fluctuating rates are deemed too unstable the policy answer is often a crawling peg. In a crawling peg regime the rate of depreciation follows a PPP path such that over time the real exchange rate remains constant (see Williamson 1965, 1982). Such a policy is an important advance over a system of occasional devaluations (too little, too late), but it is not without risks, for two reasons. First, freezing the real exchange rate may be a bad policy when disturbances in fact call for a path of, say, real depreciation. Second, there is a tradeoff between stability of the real exchange rate and price stability. A policy of fully accommodating any and all price or cost disturbances by an offsetting depreciation may in fact remove price stability altogether (see Dornbusch 1982).

PPP issues enter exchange rate policy also when a country seeks to gain macroeconomic advantages by a deliberate policy of driving the exchange rate away from PPP. A real depreciation serves to gain competitiveness and shift employment toward the depreciating country. In the 1930s this was called a 'beggar-thy- neighbour' policy, and in post-Second World War Europe it became 'export-led growth'. A policy of real appreciation, by contrast, serves to reduce inflationary pressure as the rate of increase of tradable goods prices is pushed below the prevailing rate of inflation. These macroeconomic effects of purchasing power disparities are not difficult to bring about: easy money, in the short and medium term, serves to depreciate the exchange rate and thus create employment. This policy is more effective and more lasting the more sticky wages are and the smaller the connection between wages, prices and the exchange rate is. By contrast, in an economy that is strongly indexed and in particular with exchange rate influences on indexation, an attempt at creating employment via easy money would be frustrated as exchange depreciation precipitates offsetting wage and price inflation.

Deviations from PPP have also been used as a disinflation policy. Deliberate fixing of the exchange rate or pre-announced rates of depreciation below the prevailing rates of inflation have been adopted in various countries to break inflation. The experience has been almost uniformly disappointing and worse. The resulting overvaluation very often has led to excessive external deficits, borrowing and capital flight, and ultimately only moderate success at disinflation. The cases of Chile and Argentina in the late 1970s were particularly extreme. Exchange rate policies led to extreme overvaluation. But these economies had been opened to unrestricted trade or free capital flows. The public therefore could speculate against the overvalued currency by massive imports or capital flight while the governments financed the resulting deficits by external borrowing. In the end the scheme collapsed, leaving the private sector with foreign goods or foreign assets and the governments with huge foreign debts.

PPP disparities are relevant for the exchange rate choice between flexible and fixed or managed rates. In a world where exchange rate movements conform strictly to PPP and monetary policy governs prices there is no issue. Flexible rates then allow a country to choose its preferred rate of inflation. But once disparities are possible both as a result of structural trends and perhaps as a consequence of shortterm capital movements the fixed versus flexible rate choice becomes more difficult. Flexible rates are preferable because there is no risk that the government pegs a rate that no longer corresponds to equilibrium. But flexible rates suffer the handicap that disequilibrating capital flows can drive the real exchange rate away from the level warranted by the fundamentals of the goods market. In particular, if exchange rates respond more to asset markets than price levels, persistent real appreciation or depreciation become a possibility. When this occurs there is invariably a call for PPP-based foreign exchange market intervention to bring rates back to 'fundamentals'. Explicit target zones have been proposed as a means of maintaining the advantages of flexible rates within limits to maintain approximate PPP (see Williamson 1983).

Flexible rates are also a concern because disequilibrating capital flows can provoke large changes in the rate of inflation. A loss of confidence, whether warranted or not, induces a capital outflow and a real exchange rate depreciation, as the experience of many East Asian countries in the late 1990s has demonstrated. If domestic financial policies are linked via the budget or indexation to the exchange rate, the real depreciation can initiate a sharp increase in inflation. Much of the discussion of the merits of flexible rates has concentrated on the question of whether speculative capital flows 'cause' the inflation or whether they merely respond to an inflationary situation, bringing about exchange depreciation in line with prevailing inflation. The Graham–Nurkse–Robinson view asserts, contrary to Milton Friedman, that destabilizing capital flows are the central element in the outbreak of major inflation experiences. Exchange stabilization, similarly, is seen as an essential step in stopping a runaway inflation and initiating a stabilization programme.

PPP is also relevant in the context of devaluation of a fixed rate. In the monetary approach to the balance of payments a firm tenet is the proposition that a devaluation cannot exert a lasting effect on relative prices or the balance of trade. Exchange rate depreciation raises the prices of all tradable goods in the same proportion and any effect then must be limited to a temporary depression of home goods prices due to reduced absorption. As money responds to the external surplus, real absorption rises and the initial real equilibrium is restored. This approach has the disturbing implication that devaluation does not appear to be an effective means of coping with trade or employment problems. In practice devaluation will work well when it is designed to speed up the adjustment from an initial disequilibrium in a situation where wages and prices are less than fully flexible downward. But a devaluation is likely to be ineffective if it is accompanied by a monetary expansion and wage increases, thus eliminating any real effects.

More Recent Developments

During the 1990s PPP attracted an enormous amount of interest. Presumably driven by the disbelief that such an intuitively appealing proposition about exchange rate behaviour had found little support in the data, researchers embarked in a 'search' for PPP using increasingly sophisticated time series methods. The early 1990s saw a proliferation of studies testing for PPP over the long run either by testing whether nominal exchange rates and relative prices move together (co-integrate) or by testing whether the real exchange rate has a tendency to revert to a stable equilibrium level over time (is stationary). The latter approach is motivated by the fact that the real exchange rate may be defined as the nominal exchange rate adjusted for relative national price levels and is, therefore, a measure of the deviation from PPP.

Regardless of the great interest in this area of research, the validity of long-run PPP and the properties of PPP deviations have remained the subject of an ongoing controversy. Unit root and co-integration studies generally report the absence of significant mean reversion of the real exchange rate for the recent floating experience (recent surveys include Taylor and Taylor 2004; Sarno 2005). However, the literature has been able to identify an important pitfall in studies of the long-run behaviour of PPP deviations. Specifically, one well-documented explanation for the inability to find clear-cut evidence of PPP is the low power of conventional tests to reject a false hypothesis of a unit root in (non-stationarity of) the real exchange rate (that is, the hypothesis that PPP is invalid) with a sample span corresponding to the length of the recent float (Frankel 1990). Put simply, conventional time series methods would not be able to detect the reversion of exchange rates towards PPP even if PPP were indeed valid unless very long samples of data were made available.

Researchers have sought to overcome the 'power' problem in testing for PPP in various ways. One logical reaction to tackle this problem was to test for mean reversion in the real exchange rate using long spans of data. Lothian and Taylor (1996), for example, use two centuries of data on dollar–sterling and franc–sterling real exchange rates and provide evidence supporting PPP in the recent floating period. This evidence is 'indirect' in the sense that PPP was found to hold over the full sample, which includes the recent float. Lothian and Taylor could not find any significant evidence of a structural break between the pre- and post-Bretton Woods period, and argue that the widespread failure to detect mean reversion in real exchange rates during the recent float may simply be due to the shortness of the sample.

Long-span studies have, however, been subject to some criticism in the literature. One criticism relates to the fact that, because of the very long data spans involved, various exchange rate regimes are typically spanned. Also, real shocks may have generated structural breaks or shifts in the equilibrium real exchange rate. This is a necessary evil with long-span studies of which researchers are generally aware. The long samples required to generate a reasonable level of power with standard univariate unit root tests may be unavailable for many currencies (perhaps thereby generating a 'survivorship bias' in tests on the available data) and, in any case, may potentially be inappropriate because of differences in real exchange rate behaviour both across different historical periods and across different nominal exchange rate regimes (for example, Baxter and Stockman 1989).

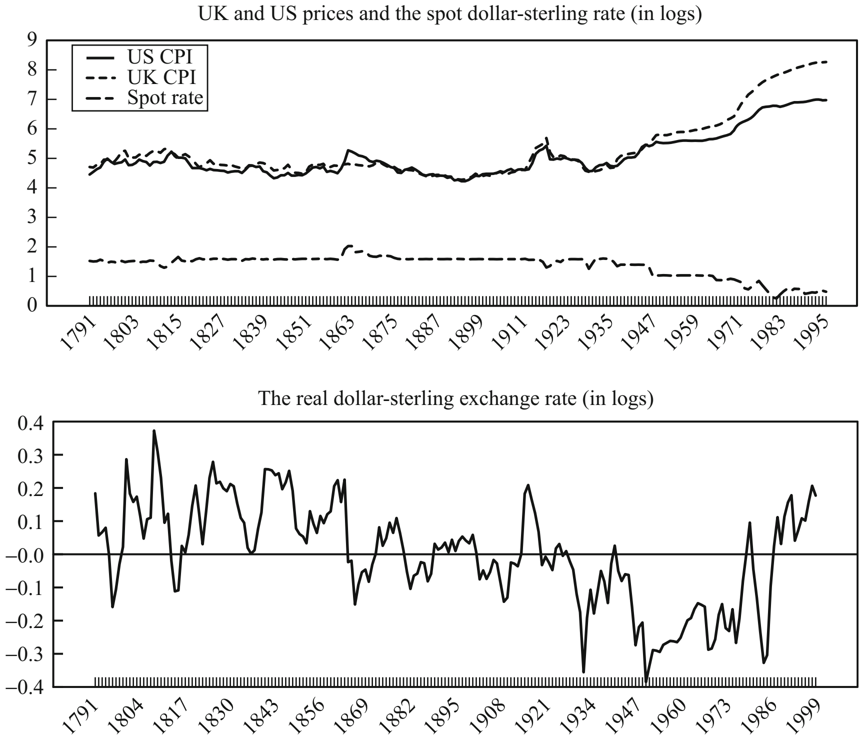
In light of the evidence provided by this literature, there remain several unresolved puzzles, among which two are prominent. First, it is still controversial whether long-run PPP is valid during the recent floating exchange rate regime from 1973 or so. Second, it is puzzling why the majority of studies which favour long-run PPP, such as the long-span studies, find empirical estimates of the persistence of PPP deviations that are too high – the half-life of shocks, that is, the time it takes for a shock to the real exchange rate to dissipate by one half, ranges between 3 and 5 years – to be explained in light of conventional nominal rigidities and to be reconciled with the large short-term volatility of real exchange rates (Rogoff 1996).

A source of potentially important bias in estimates of the half-life is caused by cross-sectional aggregation in moving from the law of one price for individual goods to PPP deviations based on price indices. Imbs et al. (2005) demonstrate how such bias is bound to be present in estimates of the real exchange half-life and provide empirical evidence that the bias is substantial. Crucini et al. (2005) adopt a similar approach to understanding the behaviour of deviations from the law of one price and PPP by examining micro-data on absolute prices of goods. They study good-by-good deviations from the law of one price for over 5000 goods and services between European Union countries for the years 1975, 1980, 1985 and 1990, and report that between most countries there are roughly as many overpriced goods as there are underpriced goods so that PPP holds to a good approximation, particularly after wealth differences are controlled for.

It is instructive to graph the real exchange rate and its components over a long span of time to speculate on its low-frequency properties. The top panel of Fig. 1 plots the time series for prices in the UK and the United States as well as the nominal dollar–sterling exchange rate over the sample period 1791–2000, with all time series expressed in logarithms.

It is quite interesting how the price series move together – even without being adjusted by the exchange rate to express prices in a common currency – over such a long period. It is also apparent how the bigger and more persistent wedge between the two prices seems to occur in the post-Bretton Woods period, essentially from the 1970s onwards. This wedge also coincided with the beginning of a corresponding trend in the nominal exchange rate, exactly as one would expect under PPP. The bottom panel of Fig. 1 then graphs the real exchange rate constructed from these time series (in deviation from the mean). The real exchange rate appears to have a tendency to return to its long-run mean (consistent with the PPP hypothesis), although the mean is crossed only 20 times in more than 200 years of data, indicating a remarkable degree of persistence in departures from PPP. Furthermore, the real exchange rate appears to be more persistent when it is in the proximity of the long-run mean, whereas reversion towards the mean happens more rapidly when the absolute size of the PPP deviation is large. This eyeball analysis of 200 years of real dollar–sterling exchange rate therefore suggests that this real exchange rate may be stationary, albeit persistent, and that it is very persistent in the neighbourhood of PPP, while being mean-reverting at a faster speed when the deviation from PPP gets larger. This is consistent with the existence of nonlinear dynamics in the real exchange rate, implying that the speed of mean reversion is state dependent.

In fact, the empirical literature on PPP has pursued formally the idea of nonlinearities in real exchange rate dynamics since the second half of the 1990s, providing several insights. In essence, the procedures conventionally applied by researchers to examine long-run PPP assume that the real exchange rate follows a linear process and depends on its own past values. In turn, this means that adjustment to the PPP equilibrium is assumed



Purchasing Power Parity, Fig. 1 200 years of prices and exchange rates (Sources: 1791–1991: Lothian and Taylor (1996); 1992–2000: International Financial Statistics database of the International Monetary Fund)

to be both continuous and of constant speed, regardless of the *size* of the past deviation from PPP. However, the presence of transactions costs may imply a nonlinear process, which has important implications for the conventional tests of long-run PPP. A number of authors have developed theoretical models of nonlinear real exchange rate adjustment arising from transactions costs in international arbitrage (for example, Dumas 1992). In most of these models proportional or ‘iceberg’ transport costs (‘iceberg’ because a fraction of goods is presumed to ‘melt’ when shipped) create a band for the real exchange rate within which the marginal cost of arbitrage exceeds the marginal benefit. Assuming instantaneous goods arbitrage at the edges of the band then typically implies that the thresholds become reflecting barriers.

Drawing on recent work on the theory of investment under uncertainty, some of these studies show that the thresholds should be interpreted more broadly than as simply reflecting shipping costs and trade barriers per se, but also as resulting from the sunk costs of international arbitrage and the resulting tendency for traders to wait for sufficiently large arbitrage opportunities to open up before entering the market.

Taylor (2001) has shown that empirical estimates of the half-life of shocks to the real exchange rate may be biased upwards because of two empirical pitfalls. The first pitfall identified by Taylor relates to temporal aggregation in the data. Using a model in which the real exchange rate follows an AR(1) process at a higher frequency than that at which the data is sampled, Taylor shows analytically that the degree of upward bias in the estimated half-life rises as the degree of temporal aggregation increases, that is, as the length of time between observed data points increases. The second pitfall highlighted by Taylor concerns the possibility of nonlinear adjustment of real exchange rates. On the basis of Monte

Carlo experiments with a nonlinear artificial data generating process, Taylor shows that there can also be substantial upward bias in the estimated half-life of adjustment from assuming linear adjustment when in fact the true adjustment process is nonlinear.

Overall, the theoretical models based on non-zero transactions costs of arbitrage in international goods markets suggest that the exchange rate will become increasingly mean-reverting with the size of the past deviation from the PPP equilibrium level. In other words, the speed at which the real exchange rate reverts to PPP depends on the size of the past deviation from PPP itself. When the real exchange rate is arbitrarily close to PPP, the real exchange rates may move randomly next period since agents have no arbitrage opportunities available in international goods markets. At the other extreme, when the real exchange rate deviates from PPP by a very large extent, it is likely that arbitrage forces will imply movements of goods and changes in prices (expressed in a common currency) that induce fast reversion to PPP.

Note that these arguments rationalize mean reversion in the real exchange rates based on ideas that relate to the law of one price in the sense that refers to tradable goods only. However, we argue that this is reasonable given that Engel (1999), in a study that measures the proportion of dollar real exchange rate movements that can be accounted for by movements in the relative prices of non-tradable goods, finds that relative prices of non-tradable goods appear to account for essentially none of the movement of dollar real exchange rates. Hence, much of the explanation for the time series properties of PPP deviations is likely to reside in the behaviour of deviations from the law of one price, that is, movements in the relative prices of tradable goods.

To turn to the empirics of nonlinear reversion to PPP, models that capture the nonlinear behaviour described above have shed light on several aspects of the behaviour of the real exchange rate. For example, Michael et al. (1997) apply a nonlinear model to monthly interwar data for the French franc–US dollar, French franc–UK sterling and UK sterling–US dollar as well as for the Lothian and Taylor (1996) long-span data-set described in Fig. 1. Their results clearly reject the linear framework in favour of a nonlinear process. The systematic pattern in the estimates of the nonlinear models provides strong evidence of mean-reverting behaviour for PPP deviations, and helps explain the mixed results of previous studies.

Using data for the recent float alone, Taylor et al. (2001) provide strong confirmation that the four major real bilateral dollar exchange rates obtaining among the G5 economies are well characterized by nonlinearly mean-reverting processes. Their estimated model implies an equilibrium level of the real exchange rate in the neighbourhood of which the behaviour of the real exchange rate is close to a random walk, becoming increasingly mean-reverting with the absolute size of the deviation from equilibrium, consistent with the theoretical literature on the nature of real exchange rate dynamics in the presence of international arbitrage costs.

Impulse response functions based on these nonlinear real exchange rate models suggest that the speed of real exchange rate adjustment is typically much faster than the very slow speeds of real exchange rate adjustment often recorded in the literature. For example, the estimated half-lives (in months) for dollar–sterling and dollar–yen are the following

Shock (%)	40	30	20	10	5	1
Dollar–sterling	10	20	22	26	29	32
Dollar–yen	14	18	24	32	38	42

where the first row reports the size of the shock (in percentage terms) to the real exchange rate. The estimated half-lives of these major real dollar exchange rates illustrate the nonlinear nature of the response to shocks, with larger shocks mean-reverting much faster than smaller shocks. The dollar–sterling rate displays quite fast mean reversion, ranging from a half-life of less than 1 year for the largest shocks of 40% to just under 3 years for small shocks of one percent; for shocks of 5–10%, the half-lives are just over 2 years. The dollar–yen displays higher persistence, with half-lives ranging from 14 to 42 months. These results therefore seem to shed some light on the PPP puzzles. Only for small shocks occurring when the real exchange rate is near its equilibrium do nonlinear models consistently yield half-lives in the range of 3–5 years, which Rogoff (1996) terms 'glacial'. For dollar–sterling, even small shocks of 1–5% have a half-life less than 3 years; for larger shocks, the speed of mean reversion is even faster.

An interesting experiment in terms of gauging the extent to which market integration and the reduction of trade costs impacts on the degree of mean reversion in real exchange rates is provided by the advent of the euro in 1999. Koedijk et al. (2004) provide empirical evidence that the introduction of the euro and, more generally, the process of economic integration in Europe have accelerated convergence to PPP, consistent with a transactions-costs goods-market arbitrage view of the mean reversion properties of the real exchange rate.

The vast empirical literature briefly reviewed here suggests that there are at least three features that are potentially important in designing a suitable model for the deviations from PPP. The first feature is that the model needs to allow for the fact that adjustment towards PPP is likely to occur at different speeds via nominal exchange rates and prices. The majority of empirical studies on PPP are based on univariate representations of the real exchange rate. This approach is valid only if certain (common factor) restrictions in the process linking exchange rates and prices are satisfied. Employing a model which does not impose these restrictions increases the power of the econometrics methods employed, while allowing us to shed light on the relative importance of nominal exchange rates and prices in restoring the PPP equilibrium. The second desirable feature is that the model allows explicitly for the possibility that different monetary and exchange rate regimes generate regime shifts in the structural dynamics of PPP deviations, especially when one uses long spans of data. The third feature is that the model might be nonlinear, in accordance with the growing evidence that exchange rate dynamics displays statistically and economically important nonlinearities.

To account for these three features, Sarno and Valente (2006) extend the long- span data used by Obstfeld and Taylor (2004) and apply a general modeling methodology in which regime changes and nonlinearities in the dynamic relationship between exchange rates and prices are explicitly allowed for, without imposing common factor restrictions. They examine the G5 countries across different exchange rate regimes, including the gold standard, the Bretton Woods period, and the floating regime since the 1970s. Over the sample period examined, the economic history of the countries involved has seen a number of fundamental changes in monetary and exchange rate regimes, institutional structure and policy targets which, in addition to the continuous evolution of the financial system and various nominal and real shocks, represent serious potential pitfalls to researchers attempting to find an empirical model of the deviations from PPP that is stable over the full sample.

Sarno and Valente's results are supportive of long-run PPP for each of the four major exchange rates examined and of a simple basic conjecture: under fixed exchange rate regimes relative prices adjust to restore deviations from long-run equilibrium, while nominal exchange rates bear most of the burden of adjustment under flexible exchange rate regimes. This is consistent with the general notion that the relative importance of exchange rates and relative prices in restoring the long-run equilibrium level of the exchange rate varies over time and is affected by the nominal exchange rate arrangement in operation. Further, the estimated half-lives of the nonlinear exchange rate models are sensibly different for fixed and floating regimes. Under fixed exchange rate regimes, shocks to the PPP equilibrium relationship may be very persistent, implying half lives – on average across the exchange rates considered – from over 5 years for large real exchange rate shocks of 20% to almost 10 years for small shocks of 1%. However, the corresponding half-lives during floating exchange rate regimes are drastically shorter, since the nominal exchange rate is allowed to operate and contribute to restoring PPP. In fact, shocks will last for less than 1 year on average for 20% shocks. The properties of PPP deviations under floating exchange regimes implied by their model appear to be fairly consistent with standard models of open economy macroeconomics and with their dynamic properties under conventional nominal rigidities (for example, Chari et al. 2002). It is only under fixed exchange rate regimes, when the burden of adjustment towards PPP relies exclusively on relative prices, that we observe remarkably long half-lives of PPP shocks.

Concluding Remarks

PPP remains an essential element of open economy macroeconomics for at least two reasons. First, it is a benchmark by which to judge the level of an exchange rate. Cassel argued that without PPP there would be no meaningful way of discussing overvaluation or undervaluation. That recognition has found a very concrete expression in the real exchange rate series now routinely calculated and reported by governments, international organizations and financial institutions. These series show exchange-rate-adjusted price relatives for a country relative to its trading partners. The series are constructed on the basis of GDP deflators, unit labour costs, manufacturing prices and wholesale prices for all major industrialized countries and increasingly for developing countries, too. They are used to judge changes in a country's external competitiveness, thus implicitly assuming, as Cassel did, that movements in *equilibrium* relative prices are negligible. Changes in real exchange rates then (and only then) unambiguously translate into changes in competitiveness from which to expect changes in trade flows and net exports. There is no question that these data provide a useful benchmark or starting point for policy discussion.

The second use of PPP is to serve as a simple prediction model for exchange rates at medium and long horizons. Under perfectly flexible wages and prices a monetary expansion would lead to equi-proportionate increases in wages, prices and the exchange rate, leaving all real variables unchanged. This combination of the quantity theory and PPP is an important insight in guiding policy. Expansionary monetary policy can be effective only if wages and prices are less than fully flexible and will be more effective the more flexible the exchange rate is. The essential channel is the real depreciation of the exchange rate that served to create employment, at least for a while. Similarly, exchange depreciation can be effective only if money wages and prices are unresponsive. Policy can be effective only if PPP fails to hold. Macroeconomic theory goes increasingly in the direction of sound microfoundations, information, contracting, and pricing models under transactions costs to explore what the basis of PPP failure is in the short run and to determine the resulting extent and persistence of policy effects.

See Also

- ► International Finance
- ► Real Exchange Rates

Bibliography

Aizenman, J. 1986. Testing deviations from purchasing power parity (PPP). *Journal of International Money and Finance* 5: 25–35.

Balassa, B. 1964. The purchasing power parity doctrine: A reappraisal. *Journal of Political Economy* 72: 584–596.

Baxter, M., and A. Stockman. 1989. Business cycles and the exchange rate regime: Some international evidence. *Journal of Monetary Economics* 23: 377–400.

Bhagwati, J. 1984. Why services are cheaper in poor countries. *Economic Journal* 94: 279–286.

Brunner, K., and A. Meltzer. n.d. *Carnegie-Rochester conference series*. Amsterdam: North-Holland.

Cassel, G. 1916. The present situation of the foreign exchanges I. *Economic Journal* 26: 62–65.

Cassel, G. 1918. Abnormal deviations in international exchanges. *Economic Journal* 28: 413–415.

Cassel, G. 1922. *Money and foreign exchange after 1914*, 1930. London: Macmillan.

Cassel, G. 1928a. *Foreign investments*, Lectures of the Harris Foundation. Chicago: University of Chicago Press.

Cassel, G. 1928b. *Post-war monetary stabilization*. New York: Columbia University Press.

Chari, V., P. Kehoe, and E. McGrattan. 2002. Can sticky price models generate persistent and volatile real exchange rates? *Review of Economic Studies* 69: 533–563.

Crucini, M., C. Telmer, and M. Zachariadis. 2005. Understanding European real exchange rates. *American Economic Review* 95: 724–738.

Dixit, A., and J. Stiglitz. 1977. Monopolistic competition and optimum product diversity. *American Economic Review* 67: 297–308.

Dornbusch, R. 1976. Expectations and exchange rate dynamics. *Journal of Political Economy* 84: 1161–1176.

Dornbusch, R. 1982. PPP exchange rate rules and macroeconomic stability. *Journal of Political Economy* 90: 158–165.

Dornbusch, R. 1987. Exchange rates and prices. *American Economic Review* 77: 93–106.

Dornbusch, R., S. Fischer, and P. Samuelson. 1977. Comparative advantage, trade and payments in a Ricardian model with a continuum of goods. *American Economic Review* 67: 823–839.

Dumas, B. 1992. Dynamic equilibrium and the real exchange rate in spatially separated world. *Review of Financial Studies* 5: 153–180.

Engel, C. 1999. Accounting for US real exchange rate changes. *Journal of Political Economy* 107: 507–538.

Fleming, M. 1962. Domestic financial policies under fixed and under floating exchange rates. *IMF Staff Papers* 9: 369–379.

Frankel, J. 1990. Zen and the art of modern macroeconomics: The search for perfect nothingness. In *Monetary policy for a volatile global economy*, ed. W. Haraf and T. Willett. Washington, DC: American Enterprise Institute.

Frenkel, J., and H. Johnson, eds. 1975. *The monetary approach to the balance of payments*. London: George Allen and Unwin.

Frenkel, J., and H. Johnson, eds. 1978. *The economics of flexible exchange rates*. Reading: Addison Wesley.

Great Britain. 1810. *Report from the Select Committee on the High Price of Gold Bullion: Ordered* by the House of Commons to be printed, 8 June 1810. Great Britain: Parliamentary Debates.

Haberler, G. 1961. *A survey of international trade theory: Special papers in international economics*. Princeton: Princeton University Press.

Harrod, R. 1939. *International economics*, 1957. Chicago: University of Chicago Press.

Holmes, J. 1967. The purchasing power parity theory: In defense of Gustav Cassel as a modern theorist. *Journal of Political Economy* 75: 686–695.

Hsieh, D. 1982. The determinants of the real exchange rate. *Journal of International Economics* 12: 355–362.

Imbs, J., H. Mumtaz, M. Ravn, and H. Rey. 2005. PPP strikes back: Aggregation and the real exchange rate. *Quarterly Journal of Economics* 120: 1–43.

Isard, P. 1977. How far can we push the 'law of one price'. *American Economic Review* 67: 942–948.

Keynes, J.M. 1923. *A tract on monetary reform*, 1971. London: Macmillan and St. Martin's Press for the Royal Economic Society.

Koedijk, K., B. Tims, and M. van Dijk. 2004. Purchasing power parity and the euro area. *Journal of International Money and Finance* 23: 1081–1107.

Kravis, I., and R. Lipsey. 1978. Price behavior in the light of balance of payments theories. *Journal of International Economics* 8: 193–246.

Kravis, I., and R. Lipsey. 1983. *Toward an explanation of national price levels*, Princeton studies in international finance. Vol. 52. Princeton: Princeton University Press.

Kravis, I., A. Heston, and R. Summers. 1978. Real GDP per capita for more than one hundred countries. *Economic Journal* 88: 215–242.

Kravis, I., A. Heston, and R. Summers. 1982. *World product and income: International comparisons of real gross product*. Baltimore: Johns Hopkins University Press.

Lothian, J., and M. Taylor. 1996. Real exchange rate behavior: The recent float from the perspective of the past two centuries. *Journal of Political Economy* 104: 488–510.

Michael, P., A. Nobay, and D. Peel. 1997. Transactions costs and nonlinear adjustment in real exchange rates: An empirical investigation. *Journal of Political Economy* 105: 862–879.

Mundell, R. 1968. *International economics*. London: Macmillan.

Mundell, R. 1971. *Monetary theory*. Pacific Palisades: Goodyear.

Mussa, M. 1979. Empirical regularities in the behavior of exchange rates and theories of the foreign exchange market. In *Policies for employment*, *prices*, *and exchange rates*, vol. 11.

Nurkse, R. 1944. *International currency experience: Lessons of the interwar period*. Geneva: League of Nations.

Obstfeld, M., and A. Taylor. 2004. *Global capital markets: Integration, crisis, and growth*. Cambridge: Cambridge University Press.

Officer, L. 1984. *Purchasing power parity and exchange rates*. Greenwich: JAI Press.

Rogoff, K. 1996. The purchasing power parity puzzle. *Journal of Economic Literature* 34: 647–668.

Samuelson, P.A. 1964. Theoretical notes on trade problems. *The Review of Economics and Statistics* 46: 145–154.

Sarno, L. 2005. Towards a solution to the puzzles in exchange rate economics: Where do we stand? *Canadian Journal of Economics* 38: 673–708.

Sarno, L., and G. Valente. 2006. Deviations from purchasing power parity under different exchange rate regimes: Do they revert and, if so, how? *Journal of Banking and Finance* 30: 3147–3169.

Schumpeter, J. 1954. *History of economic analysis*. New York: Oxford University Press.

Taylor, A. 2001. Potential pitfalls for the purchasing power parity puzzle? Sampling and specification biases in mean reversion tests of the law of one price. *Econometrica* 69: 473–498.

Taylor, A., and M. Taylor. 2004. The purchasing power parity debate. *Journal of Economic Perspectives* 18 (4): 135–158.

Taylor, M., D. Peel, and L. Sarno. 2001. Nonlinear mean-reversion in real exchange rates: Toward a solution to the purchasing power parity puzzles. *International Economic Review* 42: 1015–1042.

Viner, J. 1937. *Studies in the theory of international trade*. London: George Allen and Unwin.

Williamson, J. 1965. *The crawling peg*, Essays in international finance. Vol. 50. Princeton: Princeton University Press.

Williamson, J., ed. 1982. *The crawling peg: Past performance and future prospects*. London: Macmillan.

Williamson, J. 1983. *The exchange rate system*, Institute for international economics, policy analyses in international economics. Cambridge, MA: MIT Press.

Yeager, L. 1958. A rehabilitation of purchasing power parity. *Journal of Political Economy* 66: 516–530.

Young, J. 1925. *European currency and finance*. Commission of Gold and Silver Enquiry, United States Senate, Serial 9. New York: Garland, 1983.

Purification

Stephen Morris

Abstract

Many complete information games have equilibria only in mixed strategies, where players are required to randomize over pure strategies among which they are indifferent according to a fixed probability distribution. Harsanyi showed that, if such a game were perturbed – with each player observing an idiosyncratic independent payoff shock with continuous support – then the perturbed games will have pure strategy equilibria which will converge to the mixed strategy equilibrium as the perturbations become small. We review the strong conditions on perturbations under which Harsanyi's result holds and discuss weaker conditions on perturbations which ensure the existence of pure strategy equilibria without approximating all mixed equilibria of the unperturbed game.

Keywords

Approachability; Complete information games; Correlation; Extensive form games; Finite dynamic games; Harsanyi, J. C.; Incomplete information games; Mixed strategy equilibria; Nash equilibrium; Normal form games; Overlapping generations model; Pure strategy equilibria; Purification; Regular Nash equilibrium; Subgame perfection

JEL Classifications

C7

In a mixed strategy equilibrium of a complete information game, players randomize between their actions according to a particular probability distribution, even though they are indifferent between their actions. Two criticisms of such mixed strategy equilibria are (*a*) that players do not seem to randomize in practice, and (*b*), if a player were to randomize, why would he or she choose to do so according to probabilities that make other players indifferent between their strategies?

Since many games have no pure strategy equilibria, it is important to provide a more compelling rationale for the play of mixed strategy equilibria.

Harsanyi (1973) gave a 'purification' interpretation of mixed strategy equilibrium that resolves these criticisms. The complete information-game payoffs are intended as an approximate description of the strategic situation, but surely do not capture every consideration in the minds of the players. In particular, suppose that a player has some small private inclination to choose one action or another independent of the specified payoffs, but this information is not known to the other players. Then the behaviour of such players will look – to their opponents and to outside observers – as if they are randomizing between their actions, even though they do not experience the choice as randomization. Because of the private payoff perturbation, they will not in fact be indifferent between their actions, but will almost always be choosing a strict best response. Harsanyi's remarkable purification theorem showed that all equilibria (pure or mixed) of almost all complete information games are the limit of pure strategy equilibria of perturbed games where players have independent small shocks to payoffs.

There are other inpts of mixed strategy play: Reny and Robson (2004) present an analysis that unifies the purification interpretation with the 'classical' interpretation that players randomize because they think that there is a small chance that their mixed strategy may be observed in advance by other players. But Harsanyi's purification theorem justly provides the leading interpretation of mixed strategy equilibria among game theorists today.

I first review Harsanyi's theorem. Harsanyi's result applies to regular equilibria of complete information games with independent payoff shocks; since many equilibria of interest (especially in dynamic games) are not regular, Harsanyi's result cannot be relied upon in many economic settings of interest; I therefore briefly review what little is known about such extensions.

Harsanyi's theorem has two parts: (*a*) pure strategy equilibria always exist in suitably perturbed versions of a complete information game; and (*b*) for any regular equilibrium of a complete information game and any sequence of such perturbed games converging to the complete information game, there is a sequence of pure strategy equilibria converging to the regular equilibrium. An important literature has ignored the latter approachability question and focused on the former pure strategy existence qst, identifying conditions on an information structure – much weaker than Harsanyi's – to establish the existence of pure strategy equilibria. I conclude by reviewing these papers.

Harsanyi's th

Consider two players engaging in the symmetric coordination game below.

	A	*B*
A	2,2	0,0
B	0,0	1,1

As well as the pure strategy Nash equilibria (*A*,*A*) and (*B*,*B*), this game has a symmetric mixed strategy Nash equilibrium where each player chooses *A* with probability $\frac{1}{3}$ and *B* with probability $\frac{2}{3}$.

But suppose that, in addition to these common knowledge payoffs, each player *I* observes a payoff shock depending on the action he or she chooses. Thus,

	A	*B*
A	$2 + \varepsilon.\eta_{1A}$, $2 + \varepsilon.\eta_{2A}$	$\varepsilon.\eta_{1A}$, $\varepsilon.\eta_{2B}$
B	$\varepsilon.\eta_{1B}$, $\varepsilon.\eta_{2A}$	$1 + \varepsilon.\eta_{1B}$, $1 + \varepsilon.\eta_{2B}$

where $\varepsilon > 0$ is a commonly known parameter measuring the size of payoff shocks and (η_{1A},η_{1B}) and (η_{2A},η_{2B}) are distributed independently of each other, and player *i* observes only (η_{iA},η_{iB}). Finally, assume that, for each player *i*, $\eta_i = \eta_{iA} - \eta_{iB}$ is distributed according to a continuous density *f* on the real line with corresponding c.d.f. *F*.

This perturbed game is one with incomplete information, where a player's strategy is a function $s_i : \mathbb{R} \to \{A, B\}$. In equilibrium, each player will follow a threshold strategy of the form

$$s_i(\eta_i) = \begin{cases} A, \text{ if } \eta_i \geq z_i \\ B, \text{ if } \eta_i < z_i \end{cases}.$$

Under such a strategy, the *ex ante* probability that player *i* will choose action *B* is $\pi_I = F(z_i)$, and the probability he or she will choose *A* is $1 - \pi_i$. Thus we can re-parameterize the strategy as

$$s_i(\eta_i) = \begin{cases} A, \text{ if } \eta_i \geq F^{-1}(\pi_i) \\ B, \text{ if } \eta_i < F^{-1}(\pi_i) \end{cases}.$$

Let us look for a strategy profile (s_1,s_2) of the incomplete information game, parameterized by (π_1,π_2), that forms an equilibrium of the incomplete information game. Since player 1 thinks that player 2 will choose action *A* with probability $1 - \pi_2$ and action *B* with probability π_2, player 1's expected payoff gain from choosing action *A* over action *B* is then

$$2(1 - \pi_2) + \varepsilon.\eta_1 - \pi_2.$$

Thus player 1's best response must be to follow a threshold strategy with threshold

$$F^{-1}(\pi_i) = \frac{3\pi_2 - 2}{\varepsilon}$$

or

$$\varepsilon F^{-1}(\pi_1) = 3\pi_2 - 2.$$

Symmetrically, we have

$$\varepsilon F^{-1}(\pi_2) = 3\pi_1 - 2.$$

Thus, there will be a symmetric equilibrium where both players choose action B with probability π if and only if

$$\varepsilon F^{-1}(\pi) = 3\pi - 2.$$

For small ε, this equation has three solutions tending 0, $\frac{2}{3}$ and 1 as $\varepsilon \to 0$. These solutions correspond to the three symmetric Nash equilibria of the above complete information game, respectively: (*a*) both always choose A, (*b*) both choose B with probability $\frac{3}{2}$, and (*c*) both always choose B.

Harsanyi's purification theorem generalizes the logic of this example. If we add small independent noise to each player's payoffs, then each player will almost always have a unique best response and thus the perturbed game will have a pure strategy equilibrium. There is a system of equations that describes equilibria of the unperturbed game. If that system of equations is regular, then a small perturbation of the system of equations will result in a nearby equilibrium.

I will report a statement of Harsanyi's result due to Govindan, Reny and Robson (2003), which weakens a number of the technical conditions in the original theorem.

Consider an I player complete information game where each player i has a finite set of possible actions A_i and a payoff function $g_i : A \to \mathbb{R}$ where $A = A_1 \times \cdot \times A_I$. An equilibrium $\alpha \in \Delta(A_1) \times \cdot \times \Delta(A_I)$ is a *regular Nash equilibrium* of the complete information game if the Jacobian determinant of a continuously differentiable map characterizing equilibrium is non-zero at α (see van Damme 1991, Definition 1.5.1, p. 39).

The μ-perturbed game is then an incomplete information game where each player i privately observes a vector $\eta_i \in \mathbb{R}^{|A|}$. Player i's payoff in the incomplete information game if action profile a is chosen is then $g_i(a)+\eta_{ia}$; thus η_i is a private value shock. Each η_i is independently drawn according to a measure μ_i, where each μ_i assigns probability 0 to i's expectation of η_i being equal under any pair of i's pure strategies a_i and a_i', given any mixed strategy profile of the other players; Govindan, Reny and Robson (2003) note that this weak condition is implied by μ_i being absolutely continuous with respect to Lebesgue measure on $\mathbb{R}^{|A|}$. A pure strategy for player i in the μ-perturbation is a function $s_i : \mathbb{R}^{|A|} \to A_i$. A pure strategy profile s induces a probability distribution over actions $V_s \in \Delta(A)$, where

$$v_s(a) = \Pr_{\mu}\{\eta : s_i(\eta_i) = a_i \text{ for each } i\}$$

Theorem (Harsanyi 1973; Govindan et al. 2003) Suppose that α is a regular Nash equilibrium of the complete information game and that, for each i, μ_i^n converges to a point mass at $0 \in \mathbb{R}^{|A|}$. Then for all $\varepsilon > 0$ and all large enough n, the μ-perturbed game has a pure strategy equilibrium inducing a distribution on A that is within ε of α, that is,

$$|v_s(a) - \prod_{i=1}^{I} \alpha_i(a_i)| \leq \varepsilon$$

for all $a \in A$.

The pure strategy equilibria are 'essentially strict', that is, almost all types have a strict best response. An elegant proof in Govindan, Reny and Robson (2003) simplifies Harsanyi's original proof.

Dynamic Games

Harsanyi's theorem applies only to regular equilibria of a complete information game. Harsanyi noted that all equilibria of almost all finite complete information games are regular, where 'almost all' means with probability one under Lebesgue measure on the set of payoffs. Of course, normal form games derived from general extensive form games are not generic in this sense. This raises the question of whether mixed strategy equilibria of extensive form games are purifiable in Harsanyi's sense.

Here is an economic example suggesting why this is an important qst. Consider an infinite overlapping generations economy where agents live for two periods; the young are endowed with two units of an indivisible and perishable consumption good, and the old have no endowment. Each young agent has the option of transferring one unit of consumption to the current old agent. Each agent's utility function is strictly increasing in own consumption when young and old, and values smoothed consumption (one when young, one when old) strictly more highly than consuming the endowment (two when young, none when old). Under perfect information, this game has a 'social security' subgame perfect Nash equilibrium where each young agent transfers one unit to the old agent if and only if no young agent failed to do so in the past. But suppose instead that each young agent observes only whether the previous young agent made a transfer, and restrict attention to subgame perfect Nash equilibria. Then Bhaskar (1998) has shown that there is no pure strategy equilibrium with a positive probability of transfers (in fact, this conclusion remains true if all agents only observe history of any commonly known finite length). To see why, suppose there was such an equilibrium: if the young agent at date t does not transfer, then the young agent at date $t + 1$ must punish by not making a transfer; but the young agent at date $t + 2$ did not observe the date t outcome, and so will think that the young agent at date $t + 1$ deviated, and will therefore not make transfers; so the young agent at date $t + 1$ would have an incentive to make transfers, and not to punish as required by the equilibrium strategy.

However, Bhaskar shows that there are mixed strategy equilibria with positive transfers. In particular, there is an equilibrium where the young always transfers in the first period or if he or she observed transfers in the previous period, and randomizes between making transfers or not if he or she did not observe transfers. This strategy profile attains the efficient outcome and involves mixing off the equilibrium path only. It is natural to ask whether this equilibrium can be 'purified': suppose that each young agent obtains a small 'altruism' payoff shock that makes transfers to the old slightly attractive. The mixed strategy might then be the limit of pure strategy equilibria where the more altruistic agents make the transfers and the less altruistic agents do not. However, Bhaskar shows that the mixed strategy equilibria cannot be purified. The logic of Harsanyi's purification result breaks down because the equilibrium is not regular.

Very little is known in general about purifiability of mixed strategy equilibria in extensive form games. Results will presumably depend on the regularity of the equations characterizing equilibria and the modelling of payoff choices in the extensive form (for example, do shocks occur at the beginning of the game or at each decision node?). The best hope of a positive purification result would presumably be in finite dynamic games, where Harsanyi's regularity techniques might be applied. But Bhaskar (2000) gives an example of a simple finite extensive form game where mixed strategy equilibria are not purifiable because of the non-regularity of equilibria even for generic assignment of payoffs to terminal nodes. Mixed strategy equilibria play an important role in recent developments of the theory of repeated games. Bhaskar, Mailath and Morris (2006) report some positive and negative purification results in that context.

Purification Without Approachability

Harsanyi's purification theorem has two parts. First, the 'purification' part: all equilibria of the perturbed game are essentially pure; second, the 'approachability' part: every equilibrium of a complete information game is the limit of equilibria of such perturbed games. For the first part, Harsanyi's theorem uses the assumption of sufficiently diffuse independent payoff shocks. Only the second part required the strong regularity properties of the complete information game equilibria.

Radner and Rosenthal (1982) addressed a weaker version of the purification part of Harsanyi's theorem, asking what conditions on the information system of an incomplete information game will ensure that for every equilibrium (perhaps mixed) there exists an outcome equivalent pure strategy equilibrium. Thus they did not ask that every equilibrium be essentially pure and they did not seek to approximate mixed strategy equilibria of any unperturbed game. Each agent observing a signal with an atomless independent distribution is clearly sufficient for such a 'purification existence' result (whether or not the signal is payoff relevant). But what if there is correlation?

A simple example from Radner and Rosenthal (1982) illustrates the difficulty. Suppose that two players are playing matching pennies and each player i observes a payoff-irrelevant signal x_i, where (x_1,x_2) are uniformly distributed on $\left\{(x_1,x_2)\in\mathbb{R}_+^2 \mid 0\le x_1\le x_2\le 1\right\}$. In any equilibrium, almost all types of each player must assign probability $\frac{1}{2}$ to his or her opponent choosing each action (otherwise, that player would be able to obtain a payoff greater than his or her value in the zero sum game). Yet it is impossible to generate pure strategies of the players that make this property hold true. Another illustration of the importance of correlation for purification occurs in Carlsson and van Damme (1993), where it is shown that, while small independent noise leads to Harsanyi's purification result, small highly correlated noise leads to the selection of a unique equilibrium (the comparison is made explicitly in their Appendix B).

Radner and Rosenthal (1982) show the existence of a pure strategy equilibrium if each player observes a payoff-irrelevant signal with an atomless distribution and each player i's payoff-irrelevant signal and payoff-relevant information (which may be correlated) are independent of each other player's payoff-irrelevant signal. This result extends if players observe additional finite private signals which are also uncorrelated with others' atomless payoff-irrelevant signals. Their method of proof builds on the argument of Schmeidler (1973) showing the existence of a pure strategy equilibrium in a game with a continuum of players. Radner and Rosenthal (1982) also present a number of counter-exs – in addition to the matching pennies example above – with non-existence of pure strategy equilibrium. Milgrom and Weber (1985) show the existence of a pure strategy equilibrium if type spaces are atomless and independent conditional on a finite valued common state variable with payoff interdependence occurring only via the common state variable. Their result – which neither implies nor is implied by the Radner and Rosenthal (1982) conditions – has been used in many applications. Aumann et al. (1983) show that, if every player has a conditionally atomless distribution over others' types (that is, his or her conditional distribution has no atoms for almost every type), there exists a pure strategy ε-equilibrium. Their theorem thus covers the matching pennies example described above.

The existence of such purifications deals with one of the two criticisms of mixed strategy equilibria raised above: people do not appear to randomize. In particular, in any such purification the 'randomization' represents the uncertainty in a player's mind about how other players will act, rather than deliberate randomization. This interpretation of mixed strategies was originally emphasized by Aumann (1974).

See Also

▶ Global Games
▶ Mixed Strategy Equilibrium

Bibliography

Aumann, R. 1974. Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics* 1: 67–96.

Aumann, R., Y. Katznelson, R. Radner, R. Rosenthal, and B. Weiss. 1983. Approximate purification of mixed strategies. *Mathematics of Operations Research* 8: 327–341.

Bhaskar, V. 1998. Informational constraints and overlapping generations model: Folk and anti-folk theorems. *Review of Economic Studies* 65: 135–149.

Bhaskar, V. 2000. *The robustness of repeated game equilibria to incomplete payoff information*. Mimeo: University College London.

Bhaskar, V., G. Mailath, and S. Morris. 2006. Purification in the infinitely-repeated Prisoners' Dilemma. Discussion paper no. 1571, Cowles Foundation.

Carlsson, H., and E. van Damme. 1993. Global games and equilibrium selection. *Econometrica* 61: 989–1018.

Govindan, S., P. Reny, and A. Robson. 2003. A short proof of Harsanyi's purification th. *Games and Economic Behavior* 45: 369–374.

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

Milgrom, P., and R. Weber. 1985. Distributional strategies for games with incomplete information. *Mathematics of Operations Research* 10: 619–632.

Radner, R., and R. Rosenthal. 1982. Private information and pure-strategy equilibria. *Mathematics of Operations Research* 7: 401–409.

Reny, P., and A. Robson. 2004. Reinterpreting mixed strategy equilibria: A unification of the classical and Bayesian views. *Games and Economic Behavior* 48: 355–384.

Schmeidler, D. 1973. Equilibrium points of non-atomic games. *Journal of Statistical Physics* 7: 295–301.

van Damme, E. 1991. *The stability and perfection of nash equilibria*. New York: Springer-Verlag.